

# RECONSTRUCTING THE ACCRETION HISTORY OF THE GALACTIC STELLAR HALO FROM CHEMICAL ABUNDANCE RATIO DISTRIBUTIONS

DUANE M. LEE<sup>1</sup>, KATHRYN V. JOHNSTON<sup>2</sup>, BODHISATTVA SEN<sup>3</sup>, WILL JESSOP<sup>3</sup>

*Draft version January 29, 2015*

## ABSTRACT

Observational studies of halo stars during the last two decades have placed some limits on the quantity and nature of accreted dwarf galaxy contributions to the Milky Way stellar halo by typically utilizing stellar phase-space information to identify the most recent halo accretion events. In this study we tested the prospects of using 2-D chemical abundance ratio distributions (CARDs) found in stars of the stellar halo to determine its formation history. First, we used simulated data from eleven “MW-like” halos to generate satellite template sets of 2-D CARDs of accreted dwarf satellites which are comprised of accreted dwarfs from various mass regimes and epochs of accretion. Next, we randomly drew samples of  $\sim 10^3$ – $10^4$  mock observations of stellar chemical abundance ratios ( $[\alpha/\text{Fe}]$ ,  $[\text{Fe}/\text{H}]$ ) from those eleven halos to generate samples of the underlying densities for our CARDs to be compared to our templates in our analysis. Finally, we used the expectation-maximization algorithm to derive accretion histories in relation to the satellite template set (STS) used and the sample size. For certain STS used we typically can identify the relative mass contributions of all accreted satellites to within a factor of 2. We also find that this method is particularly sensitive to older accretion events involving low-luminous dwarfs e.g. ultra-faint dwarfs — precisely those events that are too ancient to be seen by phase-space studies of stars and too faint to be seen by high- $z$  studies of the early Universe. Since our results only exploit two chemical dimensions and near-future surveys promise to provide  $\sim 6$ – $9$  dimensions, we conclude that these new high-resolution spectroscopic surveys of the stellar halo will allow us to recover its accretion history — and the luminosity function of infalling dwarf galaxies — across cosmic time.

*Subject headings:* Galaxy: abundances — Galaxy: halo — Galaxy: stellar content — galaxies: dwarf — galaxies: early universe — stars: abundances

## 1. INTRODUCTION

The origin of the stellar halo has been a topic of intense study since the publication of the seminal paper by Eggen et al. (1962). The paper suggested that the stellar halo originated from the “monolithic collapse” of a galactic-sized primordial gas cloud. More specifically, they proposed that during this quick ( $\lesssim 100$  Myrs) collapse a very small portion of that metal-poor/free gas fragmented, due to Jeans’ instabilities, and formed stars. While the bulk of the gas would eventually form the young, metal-rich, circularly-orbiting, stellar disk of the Galaxy, these “halo” stars would instead be characterized as old, metal-poor, stars on mainly radial orbits due to the imprint of the cloud’s initial collapse. When Eggen et al. (1962) proposed this theory observations of the halo were restricted to small kinematic samples near the Sun — samples which lacked any features that might suggest that the halo was built over time via galactic mergers or accretion. However, a decade and a half later, Searle & Zinn (1978) stated in another seminal work that, in fact, some halo observations could be explained in another way. Their paper advanced the idea that differences in globular cluster abundance distribu-

tions versus galacto-centric distances in the halo were due to the “hierarchical merging” of many smaller galactic systems over the lifetime of the Galaxy. As a consequence of hierarchical merging, the stellar halo was created metal-poor because most galactic progenitors of the halo were accreted early on, which, in turn, afforded stellar inhabitants of these accreted systems little time to enrich to higher metallicities. The theory also suggested that, while less abundant, a distribution of more metal enriched stars and clusters should also inhabit the halo due to mergers over time. Consequently, it was these mergers that led to the radial orbits of stars and clusters that were earlier seen and characterized by Eggen et al. (1962).

Also bolstering the theory of hierarchical merging was the development of the theories of the formation of structures within the cold dark matter paradigm (e.g., Efstathiou et al. 1985). These theories predicted that the continuous merging of galaxies was facilitated by the parallel growth of the dark matter halos that hosted or formed the backbones of those galaxies. As a consequence, hierarchical merger formation of the stellar halo is simply a manifestation of that growth at the galactic scale.

While cosmological theory supported Searle & Zinn’s work, strong additional evidence for the theory of hierarchical merging came with the observations of halo substructure. In the early 90’s, Ibata et al. (1994) discovered the core of the Sagittarius dwarf galaxy in the outskirts of the stellar halo. Observations of this ob-

<sup>1</sup> Research Center for Galaxies and Cosmology, Shanghai Astronomical Observatory, Chinese Academy of Sciences, 80 Nandan Road, Shanghai, 200030, China; duane@shao.ac.cn

<sup>2</sup> Department of Astronomy, Columbia University, New York City, NY 10027

<sup>3</sup> Department of Statistics, Columbia University, New York City, NY 10027

viously “dying” satellite supported the assertion that stellar debris from the dwarf would follow the orbit of the accreted system. This debris would also disperse in phase-space over time and contribute to the growth of the halo. Further evidence for hierarchical merging came from the Sloan Digital Sky Survey (SDSS; York et al. 2000). This state-of-the-art project was the first global survey of the halo to extend beyond a couple of kiloparsecs from the Sun. All previous deep surveys of the halo were done in pencil beam mode — a mode where missing extended structures was virtually guaranteed. Initial results from SDSS showed a halo teeming with photometric overdensities within  $\sim 18$  kpc from the galactic center. This finding suggested that substructure was ubiquitous (Newberg et al. 2002). Majewski et al. (2003) found the tidal tails of Sagittarius wrapped around the MW by observing M-giant overdensities in the halo. The “smoking gun” for hierarchical merging came in 2006, when a clear and distinct photometric picture of the halo from SDSS revealed newly discovered dwarf galaxies and, more to the point, tidal streams (i.e. substructure) from past mergers called the “field of streams” (Belokurov et al. 2006).

The SDSS discoveries of abundant substructure in the halo led to numerous dynamical studies. Some studies determined the membership of known objects (e.g. Majewski et al. 2005) while others discovered new objects by their dynamical overdensities in phase-space (e.g. Schlafman et al. 2009). Beyond SDSS lies the next generation of galactic halo surveys. From photometry (LSST), astrometry (Gaia), and high-res abundances (APOGEE & GaLAH), we can expect to collect enough data for use in statistical analysis to actually answer some of the outstanding questions in Galactic astronomy. One outstanding question of great importance is: what is the merger history of the MW halo? With the aforementioned surveys soon at our disposal, we will have three ways of approaching this question.

A traditional photometric census of the halo (LSST) is only sensitive to mergers that are a few billion years old due to the phase-mixing of the projected phase-space dimensions of accreted structures (Sharma et al. 2010). Dynamical studies like Gaia should prove more successful in recovering accretion histories because these studies collect data that contains full 6-D phase-space information. In fact, in principle, this information allows one to calculate orbital properties (i.e., integrals of motion) for a given potential. Since the integrals of motion for a static potential are conserved, it is possible to associate debris in orbital-property space even if the halo is fully phase-mixed (Helmi & de Zeeuw 2000). However, for the outer halo (beyond 10 kpc), even Gaia cannot measure distances with sufficient accuracy, and this means that reconstructed histories of this depth (via astrometric data) are still incomplete. Furthermore, it is highly likely that rapidly occurring, violent mergers took place in the early assembly of the halo. Significant mergers of this nature should scatter normally-conserved quantities in phase-space making the extraction of merger histories from earlier epochs harder, and perhaps futile.

In the past decade, an understanding of the limitations to stellar phase-space data analysis has led of the promising pursuit of conserved quantities in stellar chemical abundance space — that is, stellar quantities which

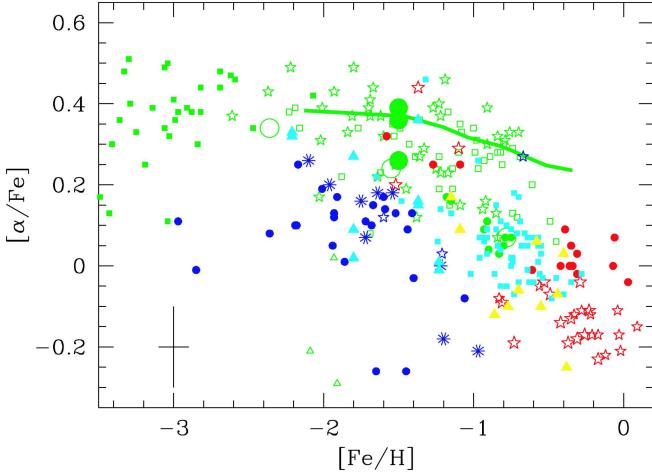
are more innate and, as such, cannot be changed by scattering in phase-space. Unavane et al. (1996) were the first to demonstrate that such innate quantities could be fruitful by using a metallicity-color ( $[\text{Fe}/\text{H}]$ -( $B-V$ )) plane to select halo stars, which are similar in composition to existing metal-poor dSph satellite stars, to constrain the hierarchical buildup of the halo. Using this comparison, Unavane determined that the history of the halo cannot contain more than  $\sim 60$  Carina-like dwarf accretions or  $\simeq 6$  Fornax-like dwarf accretions. In an analogous proposal for the Galactic disk, Freeman & Bland-Hawthorn (Freeman & Bland-Hawthorn 2002; Bland-Hawthorn & Freeman 2004) suggested that measuring the detailed chemical composition of vast numbers of the stars in the Galactic disk might be used to recover their origins: those with identical compositions in high-dimensional abundance space are likely to have been born in the same star cluster. De Silva et al. (2007) observed that star clusters are chemically homogeneous within error while Bland-Hawthorn et al. (2010) confirmed that this homogeneity allows astronomers to track stars back to the natal clusters by “chemically tagging” these stars. Thus “chemical tagging” could be used to reconstruct long-dead star clusters and recover the SFH of the Galaxy.

In this paper we explore whether a modified version of “chemical tagging” might be applied to the Galactic halo, expanding on the idea that Unavane et al. (1996) introduced over a decade earlier. Unlike stars born in the same cluster, stars born in the same dwarf galaxy do NOT share the same chemical composition. However, pioneering studies in the last decade have shown that stars in different dwarfs do have distinct (if overlapping) chemical abundance ratio distributions (CARDs; see, e.g., Nissen & Schuster 1997; Ivans et al. 1999; Shetrone et al. 2001; Venn et al. 2001; Fulbright 2002; Smecker-Hane & McWilliam 2002; Stephens & Boesgaard 2002; R. G. Gratton et al. 2003; Shetrone et al. 2003; Venn et al. 2003; Bonifacio et al. 2004; Cayrel et al. 2004; Kaufer et al. 2004; Geisler et al. 2005; Jonsell et al. 2005; Monaco et al. 2005; Johnson et al. 2006; Pompeia et al. 2006; Tautvaišienė et al. 2007). Figure 1 from Geisler et al. (2007) illustrates how these CARDs (revealed from a compilation of the aforementioned observations) tantalizingly suggests that such an attempt is possible.

Figure 1 is a reproduction of Figure 12 in Geisler et al. (2007) showing a 2-D CARDs plot of the  $[\alpha/\text{Fe}]$  (the ratio of the sum of  $\alpha$ -elements (typically, Ca, Mg, Ti, & O) to Fe) versus  $[\text{Fe}/\text{H}]$ . The plot shows various different star and star cluster measurements of  $[\alpha/\text{Fe}]$  and  $[\text{Fe}/\text{H}]$  which separate different parent or host systems into different parts of the 2-D CARD space. Additionally, differences between different galactic systems at lower metallicities are also emerging for neutron-capture elements (e.g., Strontium and Barium). These observations suggest that

- at a given accretion epoch, differences (in CARDs) between systems of the different stellar masses exist
- at a given stellar mass, differences between systems that were accreted at different times exist

In this paper, we develop a statistical approach (that



**Figure 1.** This figure is a reproduction of Figure 12 from Geisler et al. (2007). The figure is a compilation of  $[\alpha/\text{Fe}]$  vs.  $[\text{Fe}/\text{H}]$  data taken by Nissen & Schuster (1997); Ivans et al. (1999); Shetrone et al. (2001); Venn et al. (2001); Fulbright (2002); Smecker-Hane & McWilliam (2002); Stephens & Boesgaard (2002); R. G. Gratton et al. (2003); Shetrone et al. (2003); Venn et al. (2003); Bonifacio et al. (2004); Cayrel et al. (2004); Kaufer et al. (2004); Geisler et al. (2005); Jonsell et al. (2005); Monaco et al. (2005); Johnson et al. (2006); Pompeia et al. (2006); Tautvaišienė et al. (2007). Symbols shown here represent a mixture of model data, stars and star clusters found in the MW halo (green), as well as stars and stellar clusters found in low-mass dwarf spheroidals (blue), dwarf irregulars (yellow), the Sagittarius dwarf galaxy (red), and the large Magellanic Cloud (cyan). The distribution of accreted and “soon-to-accreted” systems in this 2-D chemical abundance space demonstrates the potential for determining accretion histories by attributing various subsets of the chemical abundance ratio distributions (CARDs) observed in the stellar halo of a nearby galaxy (e.g., the MW halo) to different accreted systems (see text for brief explanation).

uses the EM algorithm) to examine whether the CARDs of different mass objects accreted at different times are sufficiently different to allow us to recover halo accretion histories using data alone. We test our method with the semi-analytic models available from previous simulation work. In §2, we explain the nature of the models and methods used to produce accounts of accretion history from mock halo observations. In §3, we discuss the success of the EM algorithm when applied to specific cases. In §4, we describe the success of our results across our entire set of data. In §5, we discuss both the utility and reliability of applying this technique to real observations. In §6, we present our conclusions.

## 2. METHODS

We can approach the problem of recovering the accretion history of a galactic halo (using CARDs) by posing the following question:

“How accurately can we determine the fraction of total stellar mass,  $A_j$ , contributed by satellites of various mass ( $M_{\text{sat}}$ ) and accretion time ( $t_{\text{acc}}$ ) to a stellar halo given a set of templates for the distribution  $f_j(x_d, M_{\text{sat}}, t_{\text{acc}})$  of chemical abundances  $x_d$  found in those satellites, and observations of CARDs ( $f(x_d)$ ) in the stellar halo?”

In this study, we attempt to answer this question by investigating realizations of the stellar halo by Bullock &

Johnston (2005; see §2.1) which includes distributions of  $\alpha$ - and iron (Fe) elements generated by the methods of Robertson et al. (2005) and implemented in the models by Font et al. (2006). To begin our investigation, we define our approach by recasting our question in the form of the following equation:

$$f(x_d) = \sum_j^m A_j \cdot f_j(x_d, M_{\text{sat}}, t_{\text{acc}}) \quad (1)$$

where

$$\sum_j^m A_j = 1$$

for  $m$  satellite templates with  $A_j \geq 0$ .

In Eqn. 1,  $f(x_d)$  represents the probability density function (distribution) of observed “stars” in the  $d$ -dimensional CARD space ( $x_{1,2,3,\dots,d}$ ) and  $A_j$  represents the relative contributions from each template  $f_j$ . In a generic sense, each template  $f_j$  represents the CARD for dwarfs of some characteristic mass  $M_{\text{sat}}$  that were accreted at a characteristic time  $t_{\text{acc}}$ . Hence, finding all  $A_j$  values corresponds to recovering the “accretion history profile” (AHP) of the galactic halo. Utilizing Eqn. 1 to address our question requires the following four steps:

1. Generate mock “observations” of CARDs (i.e.  $f(x_d)$  in our case with  $[x_1, x_2] = [[\alpha/\text{Fe}], [\text{Fe}/\text{H}]]$ ) for 11 realizations from simulations of purely accretion-grown halos (§2.2).
2. Create CARD templates ( $f_j(x_d, M_{\text{sat}}, t_{\text{acc}})$ ) representing the density of stars in  $[\alpha/\text{Fe}]-[\text{Fe}/\text{H}]$  space for satellites found in selected 2-D bins of satellite mass and accretion time (§2.3).
3. Apply the expectation-maximization (EM) algorithm (a method for statistical estimation in finite mixture models [see §2.4]) to observations using satellite templates to recover their relative contribution (i.e.  $A_j$ ) to the host halo’s stellar mass (§3 and §4).
4. Evaluate the efficacy of this approach by using a summary statistic (§2.5) to encapsulate how accurate the method is in recovering the known accretion histories for each halo (e.g., see §4.2).

### 2.1. The Simulations

The simulations consist of 11 “MW-sized” halo realizations which involve a total of 1515 accreted satellites (with 100 – 150 satellites contributing to each halo) from the Bullock & Johnston (2005) work. Each dark matter host of the 11 halo realizations has a total mass of  $M_{\text{virial}}(z=0) = 1.4 \times 10^{12} M_{\odot}$  generated by merger trees using a statistical Monte Carlo method with an extended Press-Schechter (EPS) formalism (Somerville & Kolatt 1999; Lacey & Cole 1993; Bullock & Johnston 2005, and references therein). Differences in the AHP between each halo are entirely based on the randomness in the merger trees.

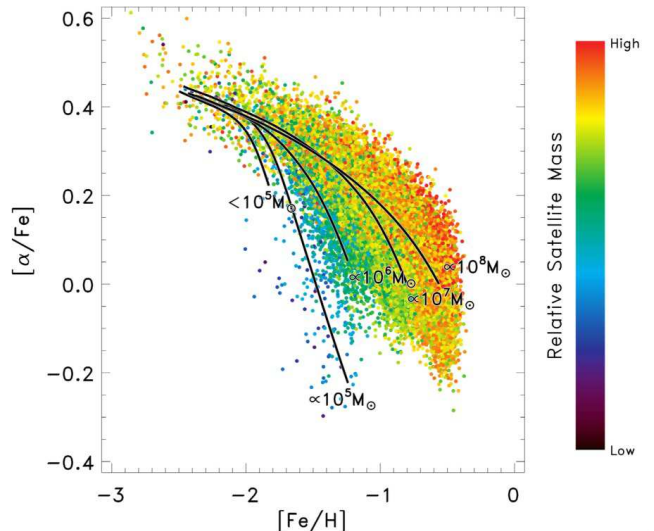
CARDs for these 11 merger histories were generated from a semi-analytic chemical enrichment code Robertson et al. (2005) which was applied separately to

each infalling satellite and combined with the simulations by Font et al. (2006). Since the enrichment code was implemented for each satellite generated, we can look at individual satellites to assess their relative contribution to their host halo’s CARDS. Also, since the aim of this study is to determine the amount of information we can retrieve via chemical abundance observations, we abstain from utilizing any of the satellites’ spatial information in our analysis. The main factors contributing to the the star formation history in the satellites are (1) the epoch of reionization,  $z_{re}$ , (2) the fraction of gas remaining/accreted in the satellite halo after reionization (set mainly by the satellite’s virial mass at its time of accretion), (3) the global star formation rate, and (4) the termination of star formation at the time of accretion (Bullock & Johnston 2005). These parameters are utilized in the simulations to determine the amount of gas available to produce stars and the duration of star formation, which, in turn, determines the chemical evolution of each satellite as prescribed in Robertson et al. (2005). The prescription includes  $\alpha$ - and Fe-element enrichment from Type II and Type Ia supernovas and stellar wind outflows of metals. The chemical evolution model was tuned with a supernova (SN) feedback treatment to agree with the local dwarf galaxy stellar mass-metallicity (Robertson et al. 2005, ; see §2.3 for further discussion). The  $\alpha$ -element patterns in dwarfs versus the smooth halo are consistent with the CARDS of dwarfs found in the compilation of data in Figure 12 of Geisler et al. (2007) (see Figure 1) — an agreement that further bolsters our approach in this investigation (Font et al. 2006).

### 2.2. “Observations” from the Simulations

The function  $f(x_d)$  represents the density distribution produced by  $n$  random “observations” in chemical abundance space  $x_d$  of “stars” (star particles; see §2.1 for explanation). Sample distributions for each halo are constructed by randomly drawing “stars”. To mimic observational errors during mock observations, we add a random number drawn from a Gaussian with a dispersion of 0.05 dex to both  $[\alpha/\text{Fe}]$  and  $[\text{Fe}/\text{H}]$  abundance ratios. The choice of the size of these errors is meant to probe the foreseeable potential of this technique by employing the best possible conditions for analysis. Evaluation of this technique with ideal conditions provides us with a baseline for expectations from which analysis of real observations in the future can be assessed. In our study, we select samples of  $nd \approx 10^3$ ,  $10^4$ , and  $3 \times 10^4$  representing current, near-future, and optimistically-anticipated sample sizes, respectively (Ken Freeman, private communication).

Figure 2 shows a 2-D CARD ( $[\alpha/\text{Fe}]$  vs.  $[\text{Fe}/\text{H}]$ ) of  $\sim 3 \times 10^4$  star particles representing mock stellar abundance ratio observations from the halo 1 simulation. The color of each particle represents the stellar mass of its parent satellite relative to all other accreted satellites. Black and purple particles are donated from the least massive satellites while orange and red particles are donated from the most massive satellites. The distribution of particles shown demonstrate the expectation that the most massive satellites should account for the vast majority of stars found in the accreted halo stellar population. In comparing this 2-D CARD to the observed CARDS in



**Figure 2.** Plot of  $[\alpha/\text{Fe}]$  vs.  $[\text{Fe}/\text{H}]$  for  $\sim 3 \times 10^4$  “star particles.” Each particle is color-coded to represent the relative stellar mass/luminosity of its parent satellite. The relative number of particles in the accreted satellite mass/luminosity range reflects the expected relative contribution from each parent to the total stellar mass of the host halo. The chemical evolution tracks of five satellites, randomly chosen to span the stellar mass range of accreted satellites for halo 1, are displayed over the colored particle distribution as black lines and labeled by a stellar mass proportional to the typical satellite stellar mass found in the mass bins outlined in §2.3 and displayed in Figure 3.

Figure 1, we see that the distributional spread between observed accreted dwarfs of different masses mirror the distributional spread (in mass) for the simulated dwarfs.

The black dashed lines that overlay the colored particle distribution of Figure 2 represent chemical evolution tracks (from the simulations) of typical dwarf masses accreted over the lifetime of the halo. The length of these tracks are primarily affected by the satellite’s accretion time. The more time a satellite has to produce stars, the longer its galactic chemical evolution can continue to advance to higher metallicities, and vice versa. The curvature of these tracks is primarily determined by the satellite’s mass. The more mass a satellite has to produce stars, the higher its star formation rate (SFR), which means chemical enrichment by core-collapse SN is greater. This enhanced early enrichment from core-collapse SN leads to higher galactic metallicities before the typical 1 Gyr onset (delay) in Type Ia SN begins (ends) to establish a so-called  $[\alpha/\text{Fe}]$ -knee via significant contributions to Fe abundances. The incorporation of these various tracks into our dwarf model templates are discussed in the next section.

### 2.3. Satellite Template Sets

To see if we can recover the AHP of our simulated halos from our mock observations we need to generate templates which represent typical accretion events of given satellite stellar mass and age. The most “naive” approach to creating our templates would be to evenly divide the possible range in time  $t_{acc}$  (0–13 Gyrs) and mass (stellar)  $M_{sat}$  ( $10^{-9} M_{\odot}$ ). This division would form  $N_r$  mass-binned templates (rows) by  $N_c$  time-binned templates (columns) with some “empty” templates ( $N_{empty}$ ) where the total number of templates equal  $N_{temps} = N_r \times N_c - N_{empty}$ . However, since decades in galactic

(stellar) mass have intuitive implications for galaxy evolution, we restrict our current templates to even divisions in  $t_{acc}$  while we divide  $M_{sat}$  by decades of mass from  $10^5 M_\odot$  to  $10^9 M_\odot$  and combine all satellites below  $10^5 M_\odot$  into a  $5^{th}$  mass bin (see Figure 3).

After divisions in the  $t_{acc} - M_{sat}$  plane are selected, all 1515 dwarf satellite models are divided amongst the bins created by the selected partitions based on each dwarf's individual  $t_{acc}$  and  $M_{sat}$ . During the process, each dwarf's chemical track (see §2.2) is smeared out by a convolution of each star particle with an observational error of  $\sigma_{err} = 0.05$  dex in both chemical dimensions. To generate the CARDS required for implementation of our recovery algorithm (i.e. the EM algorithm), we separate an average of  $\sim 19,500$  star particles per satellite (with errors) into square bins of 0.1 dex that span 3 dex in  $[\text{Fe}/\text{H}]$  ( $-3 - 0$  dex) and 1.7 dex in  $[\alpha/\text{Fe}]$  ( $-0.7 - 1$ ). The collection of all binned distributions in our 2-D chemical space are normalized to produce an ensemble of probability densities that represent our satellite template set (STS).

Figure 3 shows our 5x5 STS as an example of our model template scheme. The full 5x5 panel (top-right) shows the evenly-spaced bins in accretion time versus bins spaced out by decades of accreted satellite stellar mass down to  $10^5 M_\odot$ , below which all other satellites are binned together. As stated in §2.1, the feedback prescriptions in the chemical evolution models were tuned to reproduce the chemical abundance relationships observed in galactic surveys. First, the mass (luminosity) versus metallicity ( $[\text{Fe}/\text{H}]$ ) relationship can be seen by inspecting the trends along any accretion time column. This relationship shows an increase in the distribution peak value of  $[\text{Fe}/\text{H}]$  (and a decrease in the distribution peak value of  $[\alpha/\text{Fe}]$ ) with increasing mass (luminosity) of the galaxy. Second, an age-metallicity relationship can be seen by inspecting the trends along any accreted satellite mass row (i.e. when holding the mass range constant). This relationship shows an decrease in the distribution peak value of  $[\text{Fe}/\text{H}]$  (and a increase in the distribution peak value of  $[\alpha/\text{Fe}]$ ) with an increase in the accretion time epoch (i.e. which dictates the available time for star formation) of the galaxy. Although, it should be noted that this age-metallicity relationship is not strictly expected to hold for any given set of dwarf galaxies as other processes are as likely to quench star formation in dwarfs before accretion takes place.

Projections of the 5x5 STS, in accreted satellite mass and accretion time, are shown in top-left and bottom-right corners of Figure 3, respectively. A comparison of both projections reveals smaller differences in CARDS between adjacent bins of accretion time than in adjacent bins of accreted satellite mass. The similarities between dwarf models in the 5x1 STS projection suggests that the EM algorithm will perform better when utilizing the 1x5 STS projections of accreted satellite mass for estimates (see §3.2, §3.3 & §4.2 for further discussion). Finally, a 1x1 STS projection displaying the probability density function of our master template (i.e. containing the CARDS of all 1515 simulated dwarfs) is shown in the bottom-left of the figure.

In §3, we use the two 1-D projections discussed here to form a basis of analysis for the EM algorithm's per-

formance and our ability to recover AHP of halos in one dimension of mass or time.

#### 2.4. Recovering AHPs using the EM Algorithm

The composition of our halos can be best described as a finite mixture of discrete accreted objects that exhibit varying characteristics in a shared CARD space ( $x=[\alpha/\text{Fe}], y=[\text{Fe}/\text{H}]$ ). Since we can construct models for these accreted objects, we can create a mixture model

$$f(x_i, y_i) = \sum_{j=1}^m A_j f_j(x_i, y_i) \quad (2)$$

where the relations

$$\sum_{j=1}^m A_j = 1 \quad \text{for } A_j \geq 0, \quad j = \{1, \dots, m\}$$

confine the relative contribution of model satellites  $\mathbf{A}$ . Given  $n$  observations of  $\{x_i, y_i\}$ , we can construct a log-likelihood function as follows

$$\begin{aligned} L(\mathbf{A}) &= \prod_{i=1}^n f(x_i, y_i) \\ &= \prod_{i=1}^n \left\{ \sum_{j=1}^m A_j f_j(x_i, y_i) \right\} \\ \log L(\mathbf{A}) &= \sum_{i=1}^n \log \left( \sum_{j=1}^m A_j f_j(x_i, y_i) \right) \end{aligned} \quad (3)$$

Maximizing  $\log L(\mathbf{A})$  will yield the maximum likelihood estimate  $\mathbf{A}_{MLE}$  for  $\mathbf{A}_{EM}$  — our best expectation-maximization estimate for the *true*  $A_j$  values  $\mathbf{A}_T$ . This task, which can be computationally arduous, can be made tractable by adding a latent indicator,  $z$ , to each observed data point  $(x, y)$ , to represent the model template of origin. By designating data set  $\{x_i, y_i, z_i\}_{i=1}^n$  as our complete data, we can then define a complete data likelihood as

$$L(\mathbf{A}) = \prod_{i=1}^n \prod_{j=1}^m \left\{ A_j f_j(x_i, y_i) \right\}^{z_{ij}} \quad (4)$$

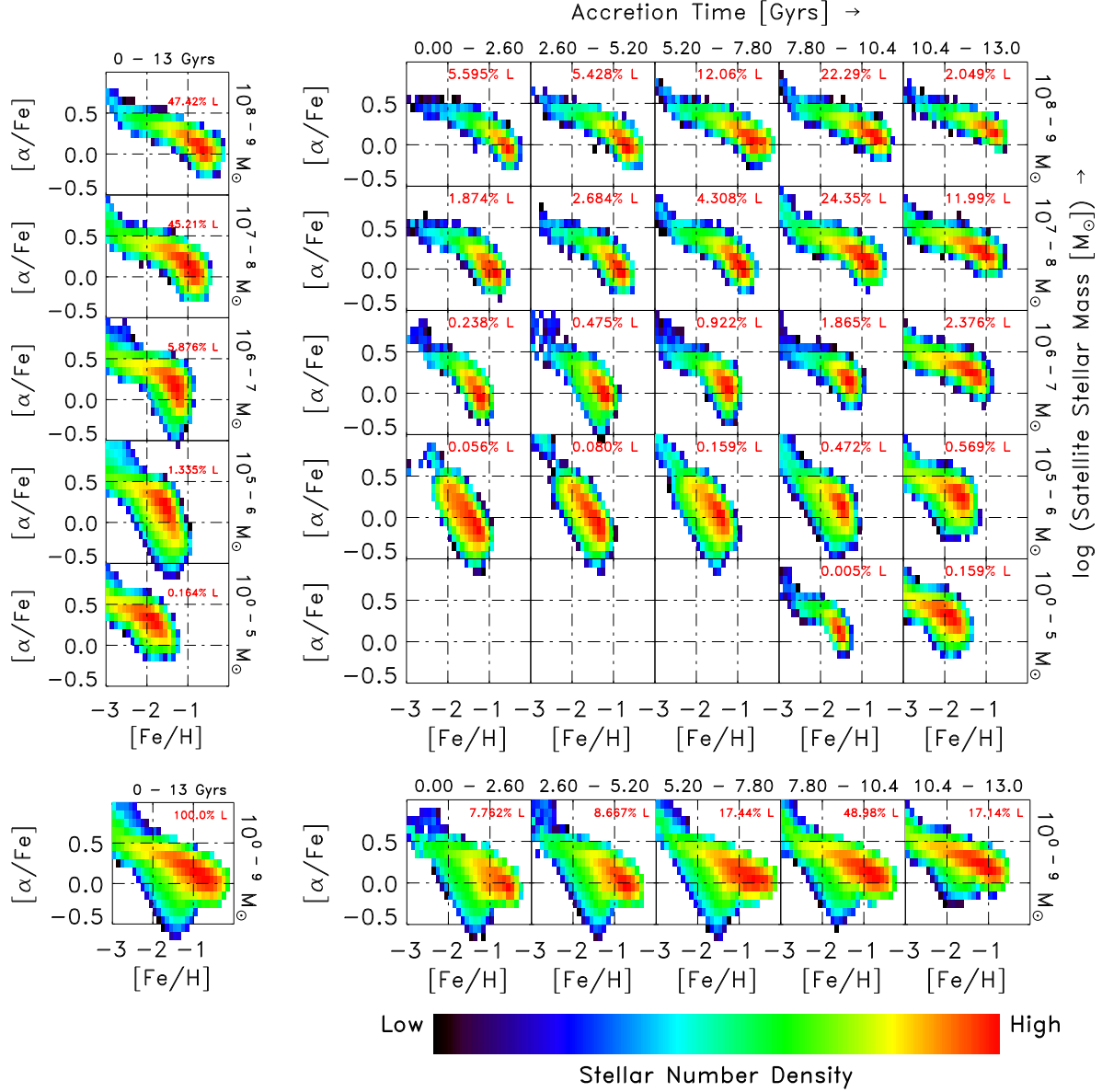
$$\ell(\mathbf{A}) = \sum_{i=1}^n \sum_{j=1}^m z_{ij} \log \{ A_j f_j(x_i, y_i) \} \quad (5)$$

where  $z_{ij}$  equals the *hard* expectation that  $(x_i, y_i)$  comes from  $j^{th}$  satellite template and  $\ell(\mathbf{A})$  is the complete data log-likelihood.

As stated above, the log-likelihood derived above can be used to obtain  $\mathbf{A}_{EM}$  via the expectation-maximization (EM) algorithm. Starting from an initial set of guesses,  $\mathbf{A}^{(0)}$ , the algorithm iteratively steps through guesses (which are informed by the former set) until the value of the log-likelihood  $\ell(\mathbf{A})$ , conditioned on the data (and within some tolerance), is maximized. More specifically, the maximizing value of the  $t^{th}$  iteration,  $\mathbf{A}^{(t)}$ , is then used as the starting value for the next run, and it continue until the likelihood changes by less than  $10^{-3}$  over twenty-five iterations. Details to the implementation of this technique are shown in Appendix A.



# "Naive" 5x5 Satellite Template Set



**Figure 3.** Plot of 5x5 STS along with projects in the  $t_{acc} - M_{sat}$  plane. *Top-right:* Our 5x5 STS. The relative contribution of stellar mass from a subset of all 1515 satellites in each template is shown as percentages of the total halo stellar mass (red). Each column and row reflects the mass/stellar mass-metallicity relation and age-metallicity relation, respectively (see §2.1 for details). *Top-left and bottom-right:* Projections of the 5x5 STS into the  $t_{acc}$  plane (top-left) and  $M_{sat}$  plane (bottom-right) are equivalent to the 1x5 (mass-divided) STS and 5x1 (time-divided) STS explored in §3. *Bottom-left:* Plot of a projection into both parameter dimensions exemplifies a density distribution (i.e.  $F(x_d)$ ) similar to the parent distributions of individual halos from which “observed” stars are drawn in our analysis.

We discuss how we evaluate the success of our estimates in the next section. Results from the EM estimates are discussed from §3 onwards.

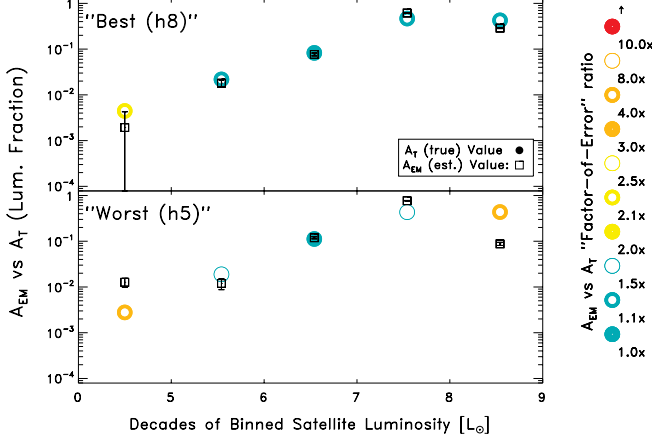
## 2.5. Evaluating the success of the method

In order to evaluate the relative success among our calculated AHPs across all halos and the success of the technique across various STS, we compare the EM estimates,  $\mathbf{A}_{EM}$ , to the known true values,  $\mathbf{A}_T$ . Using these values we can calculate the “factor-of-error” (FoE) ratio for each template EM estimate. The FoE value is defined

as the maximum between  $A_{EM,j}/A_{T,j}$  and  $A_{T,j}/A_{EM,j}$ .<sup>4</sup>

One way to evaluate the fidelity of our results is to determine an average FoE ratio ( $\langle \text{FoE} \rangle$ ) from all FoE measured (i.e. from a given STS and halo). This  $\langle \text{FoE} \rangle$

<sup>4</sup> This definition is chosen to obtain the most general sense of FoE statements (which are common in astronomy) such as “the observed [generic] measurements are within a factor of 2 of theoretical predictions.” This statement implies that observed measurements are between less-than-twice and greater-than-half of the theoretical values in question.



**Figure 4.** A plot of fractional stellar mass contributions to the host halo versus the satellite’s binned stellar mass for the *best* and *worst* EM estimates among our 11 halos (labelled h1–h11, hereafter) for 1x5 STS. Selection of these halo estimates are based on their (FoE) values, given in respect to the number of stars (here we use  $\sim 10^4$  stars) observed. Estimates from observations (open squares) are shown for each of the five templates. Their corresponding actual values (circles) are also shown with various holes and colors that indicate the “factor-of-error” difference between the estimate and actual values (see legend for key). See text for discussion.

is an average of all  $\text{FoE}_j$ , weighted by  $w_j$ , and given as

$$\langle \text{FoE} \rangle = \sum_{j=1}^m w_j \cdot \text{FoE}_j \quad (6)$$

where  $w_j$  represents a choice of weights for the relative importance of each template estimate and  $m$  is the number of templates used. The lowest  $\langle \text{FoE} \rangle$  value indicates the best results balanced by  $w_j$  in STS templates for each halo examined. For our primary analysis we take a mean of FoE values ( $w_j = m^{-1}$ ) while other weights are examined in §5. The method of evaluation is applied to results in §3–§4.1.2.

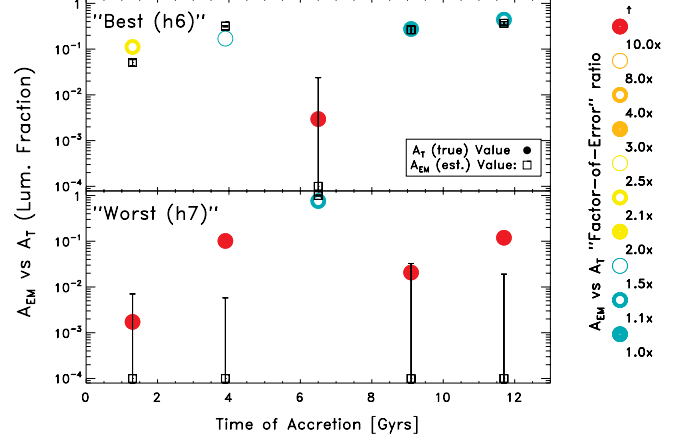
### 3. RESULTS I: ACCRETION HISTORY PROFILES IN 1-D

In this section, we determine how accurate our satellite contribution estimates can be for our simplest STS. More explicitly, we investigate how well we can estimate the fractional contributions to a halo’s construction via STS that span the stellar mass of the accreted system (i.e., its luminosity function) or its time of accretion (i.e., its stellar mass accretion history).

#### 3.1. Stellar mass fractions

As discussed in §2.3, we can construct a *true* AHP from our model stellar halos to determine how accurately we can estimate them using the EM algorithm discussed in §2.4. Here, we examine the accuracy of our 1x5 STS estimates which are a 1-D set of 5 mass bins (as described in §2.3 and shown in the top-left of Figure 3) — that is to say, we evaluate how well we can recover stellar mass fraction contributions from satellites with no sensitivity to their time of accretion.

Figure 4 presents some characteristic results from our 1x5 STS analysis. The top panel legend indicates that open squares represent the  $\mathbf{A}_{EM}$  values estimated by applying our EM analysis to observed abundances from



**Figure 5.** Figure is the same as Figure 4 for 5x1 STS. See text for discussion.

$\sim 10^4$  observed stars.<sup>5</sup> Error bars (calculated from the Fisher information matrix) indicate the smallest possible ( $1\sigma$ ) error values (see Appendix A for details). The colored circles shown represent the  $\mathbf{A}_T$  (true) values while the specific colors of each circle categorize the FoE between  $\mathbf{A}_{EM}$  and  $\mathbf{A}_T$  values by the color legend to the right of the plot. Various FoE values spanning less than 1.1 (green “solid” circle) to 10 or more (red “solid” circle) are examined.

In the figure, two plots are chosen to display results from two representative halos (labelled by “h” with the designated number for the halo for short). The two halos are the *best* (h8) and *worst* (h5) AHP estimates as determined by their average FoE ( $\langle \text{FoE} \rangle$ ) values.

Looking at our *best* EM estimates from h8, we see that individual  $\mathbf{A}_{EM}$  produce errors that are within a factor of 2.5 or better for all template estimates using  $\sim 10^4$  observed stars. This remarkable considering that we are characterizing  $\lesssim 10^{-2}$  to  $10^{-3}$  of the total halo luminosity for the lowest mass bins.

Our *worst* EM estimates from h5 seems to reinforce the notion that this analysis provides reliable results. In this worse case scenario, most estimates are within a factor of 2 while the worse estimate (given for our most massive satellite template) is within a factor of 8.

#### 3.2. Accretion time histories

The other principle dimension of our analysis is time. Using the same prescribed analysis above we can examine the success of estimating AHP from a 1-D set of 5 equally-spaced time bins (also described in §2.3) — that is to say, we evaluate how well we can recover stellar mass fraction contributions from satellites with no sensitivity to their stellar masses. Figure 5 presents some characteristic results from our 5x1 STS analysis. In the figure, plots are chosen based on the same criteria used in making Figure 4. The *best* EM estimates from h6 reveal very different results concerning the reliability of our analysis

<sup>5</sup> In a similar effort to this work, Schlafman et al. (2012) analyzed the  $[\text{Fe}/\text{H}]$  and  $[\alpha/\text{Fe}]$  chemical signatures of 9005 SEGUE stars in the MW (smooth) halo to ascertain the relative contributions to the accreted structure of the smooth halo finding a strong correlation between the SEGUE data and the accretion formation of MW halo analogs in  $N$ -body simulations at distances beyond 15 kpc from the Galactic center. Our choice of sample size demonstrates another way in which this dataset might be used.

when compared to the 1-D *mass-resolved* templates results. While both the two most recent and two earliest accretion events have FoE values  $\leq 2.5$ , the “medieval” accretion event has a FoE value  $\gtrsim 30$ . Here, only the least massive accretion event has a poor FoE value.

Our *worst* EM estimates from h7 follow a trend where all but the most massive accretion event (the medieval event in this case) have markedly poor FoE values that range from  $\gtrsim 20$  to  $\gtrsim 10^3$ . Here, the best estimate has a FoE  $\leq 1.5$  (i.e. with 50% of the true value). While the estimates call into question the reliability of using multiple dimensions in  $t_{acc}$  and  $M_{sat}$ , the overall results were already anticipated from the visual inspection of these templates in the bottom-right corner of Figure 3. As suggested earlier, it is likely that degeneracies in CARDs within this template set led to the poor  $\mathbf{A}_{EM}$  estimates seen. In particular, the difference between FoE values for the medieval accretion events in h6 and h7 versus the other events comes down to the dominant accretion event templates subsuming those events that are both highly degenerate in CARD space and significantly less massive than the main event(s). As a consequence, it may appear hopeless to try to glean any information about the accretion times from 1-D estimates. This may also hold true for estimates in multiple dimensions when accretion time is treated as the dominant dimension of analysis (see §4 for further discussion).

### 3.3. Accuracy of stellar mass fractions across halo realizations: $\langle \text{FoE} \rangle$

Our complete results, summarized by  $\langle \text{FoE} \rangle$ , provide us with insights into the overall effectiveness of the analysis for all 11 halos. Figure 6 displays  $\langle \text{FoE} \rangle$  values for the 1x5 STS (i.e. 1-D *mass-resolved*; top panel) and the 5x1 STS (i.e. 1-D *time-resolved*; bottom panel). In both panels, each plot shows an histogram of  $\langle \text{FoE} \rangle$  values, calculated using the number of observed stars indicated in each plot, and normalized by the number of halos examined. Dotted, light-grey lines indicate a  $\langle \text{FoE} \rangle = 2$  which indicates, by eye, the vast difference in trying to recover AHPs from 1-D mass of accreted satellites templates versus 1-D time of accretion templates.

In our *mass-resolved* (1x5 STS) estimates (top panel), we can examine the overall success of these templates and note the degree of improvement in estimates as a result of using more data points. Looking at the full panel, we can clearly see the gradual, distinct improvement in  $\mathbf{A}_{EM}$  estimates when a larger dataset is used. The median  $\langle \text{FoE} \rangle$  (i.e., our accuracy) for each larger set of observed stars are  $\sim 2.55$ ,  $\sim 2.16$ , and  $\sim 2.06$ , respectively. However, it is important to note that the modest improvement between the last two datasets possibly indicates that the method is hitting a limit due to number of templates versus the numbers of stars used. In our *time-resolved* (5x1 STS) estimates (bottom panel), we can see that these estimates are far poorer than the estimates for the *mass-resolved* estimates. In fact, the *time-resolved* estimates have a median  $\langle \text{FoE} \rangle$  for each larger set of observed stars equal to  $\sim 100$ ,  $\sim 175$ , and  $\sim 192$ , respectively.

Even more critical is the fact that these estimates get marginally worse with number of observations used. This suggests that there are degeneracies between templates in the set that cannot be removed with more CARD information in just two chemical abundance ratio dimen-

sions. Conversely, these degeneracies may also suggest the inherent need for mass divisions in the STS to see differences in templates — a possibility that motivates moving our STS to higher dimensions in the  $t_{acc} - M_{sat}$  plane. In the next sections, we discuss the impact of expanding our analysis to multiple dimensions in order to achieve better estimates.

## 4. RESULTS II: ACCRETION HISTORY PROFILES IN 2-D

Now that we have established a baseline for estimates in our special 1-D cases, we seek to extend our search in higher dimensions of time (i.e. fixing 5 mass bins and varying our number of equally-spaced time bins). In the following subsections, we discuss our results in detail for our 2x5 and 3x5 STS (i.e. with 2 or 3 time bins), presenting insights into their success and failure.

### 4.1. Xx5 satellite template set results

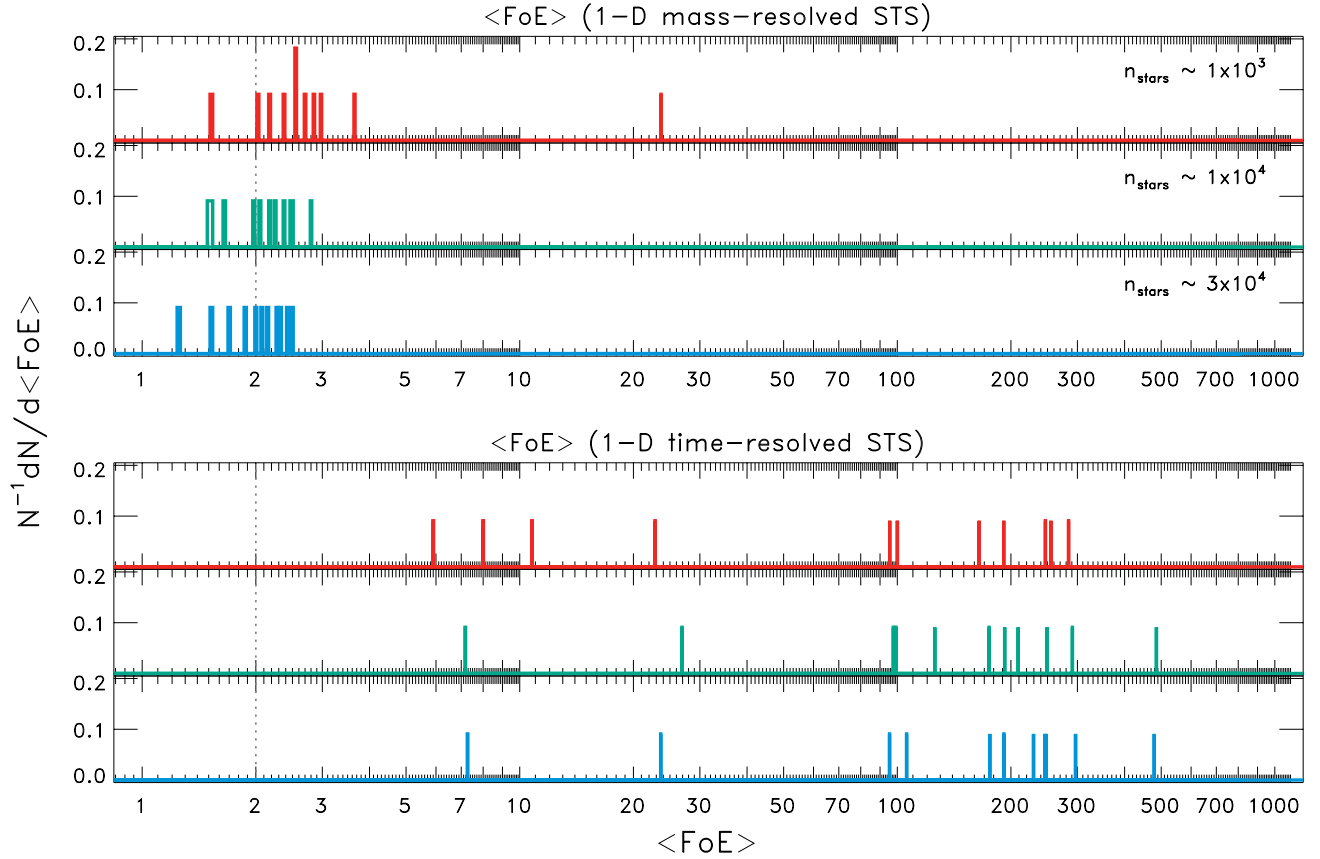
The goal of expanding our STS into higher dimensions is two-fold. First, we want to directly recover AHPs with high fidelity by dividing our  $t_{acc} - M_{sat}$  plane into templates that would reveal interesting information (e.g., about the MW halo’s history) when applied to real abundance observations. Second, we want to indirectly recover 1-D stellar mass functions (mass-resolved profiles) and time of accretion histories (time-resolved profiles) of our halos by summing “like” estimates in time or mass together (marginalization) in order to generate better accounts in 1-D than could be done directly. Our hypothesis is that allowing a finer grid in time will produce templates with less degeneracy and allow a better recovery of AHP. Of course, this must be balanced by the size of our sample and its ability to constrain the additional parameters (larger  $\mathbf{A}_{EM}$  set) from the increased number of templates.

#### 4.1.1. “Early” vs. “recent” accretion: 2x5 STS results

To address our goals, we start by generating templates for our 2x5 STS which have two, evenly divided, time bins for *recent* (0–6.5 Gyrs ago) and *early* (6.5–13 Gyrs ago) epochs. Figure 7 displays a selection of results that reveal the success of EM estimates due to the application of our 2x5 STS. In the figure, we can once again examine the *best* (h11), the *median* (h2), and the *worst* (h7) of the halo estimates using these templates. Here, in the first column of Figure 7, we display the values of  $\langle \text{FoE} \rangle$  to indicate the success of estimates using  $\sim 10^4$  stars which can be compared to the our *marginalized* results in the right-most columns. At first glance, we see that all panels indicate, by (mostly green) colors, that most estimates are within a FoE of 2. For the *best* EM estimates (from h11), it is encouraging that all FoE values are  $\leq 2$ .

However, for the *worst* EM estimates (from h7) we see a marked decrease in the fidelity of a couple of estimates and especially for one at the high mass end. Here, we see that the  $A_{r,j}$  value for the *early* accreted  $10^{7-8} M_{\odot}$  template is actually similar to its recently accreted counterpart whereas the EM estimates are very different. While the *early* accretion event is estimated to be essentially non-existent, both the adjacent higher mass template (*early* accreted  $10^{8-9} M_{\odot}$  template) and the *recent* accretion  $10^{7-8} M_{\odot}$  template have slightly higher EM estimates than their true values. The  $\leq 50\%$





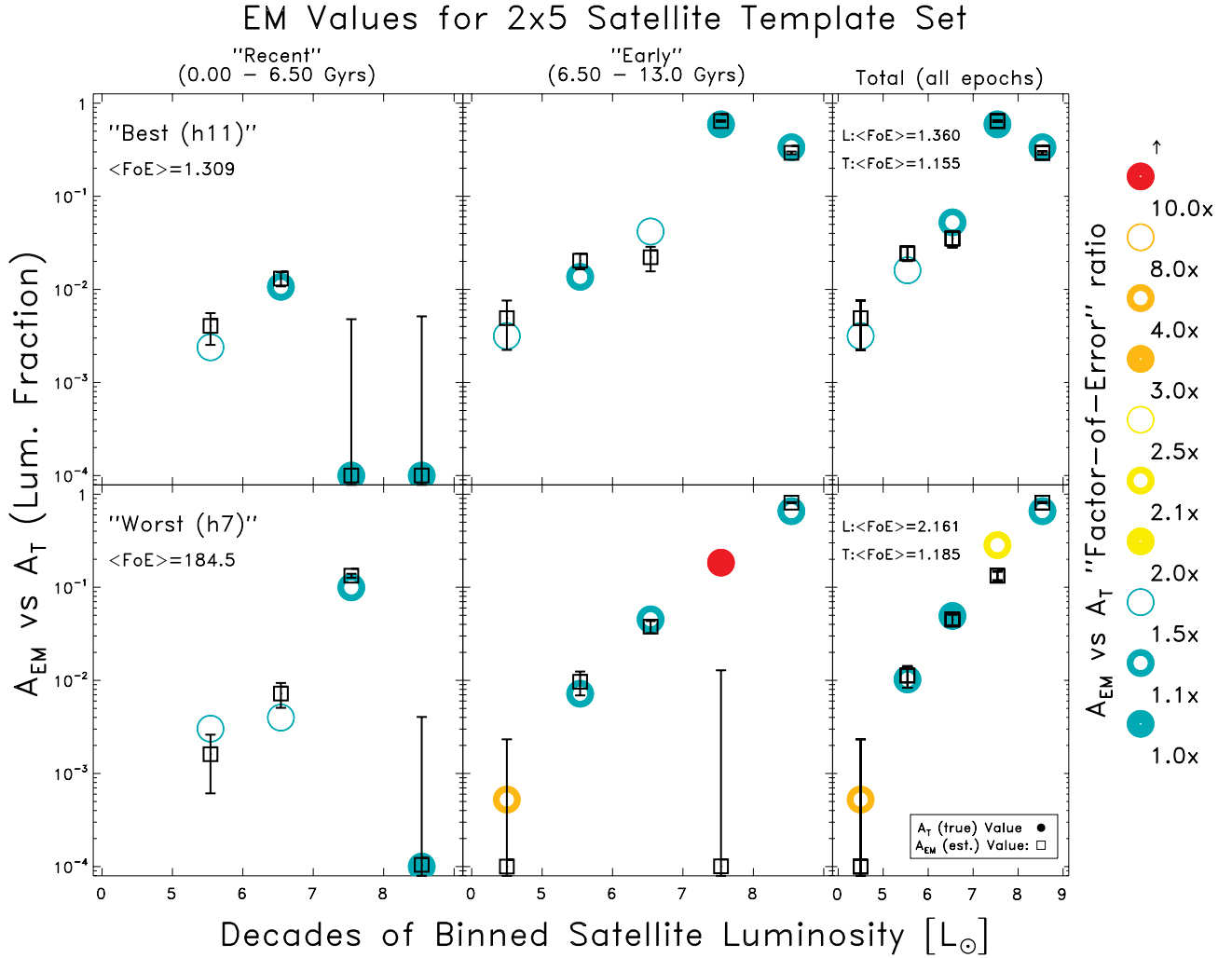
**Figure 6.** Figure shows panels for the frequency of  $\langle \text{FoE} \rangle$  values amongst all 11 halo for 1x5 STS (i.e. versus mass of accreted satellite; *top*) and 5x1 STS (i.e. versus time of accretion; *bottom*). Red, green, and blue histograms refer to the number of stars used to calculate the EM estimates summarized in this figure. Light-grey dotted lines indicate a  $\langle \text{FoE} \rangle = 2$  to guide the eye when comparing the difference in results. The difference in the spread and range of  $\langle \text{FoE} \rangle$  values between the 1x5 vs 5x1 STS are striking and seem to support the notion (from Figure 3) that 1x5 STS retains greater distinction between its templates than the 5x1 STS do (resulting in better estimates from the 1x5 STS).

difference in FoE values is probably due to both templates subsuming the contributions from the poorly estimated  $10^{8-9} M_{\odot}$  template. Given that this template is high mass and accreted early, this degeneracy is likely due to the fact that the accretion of most massive systems happens early in most of the 11 halos' histories. Since the 1515 satellites used to make the templates are comprised of 11 ensembles of accreted dwarf systems that make up the composition of our simulated halos, it is not surprising that a coarse divisions in accretion epochs lead to disparities in the fidelity of our estimates across the 6.5 Gyr divide.

On the other hand, as indicated by our *best* selection, it is reassuring that given the simplicity of our dwarf models, there is enough information in their CARDS to make templates that differentiate between higher mass progenitors of the halo at different epochs. This is true, despite the fact that the highest mass dwarf models show the greatest amount of degeneracy among accreted systems throughout all halos' assembly histories. Also, given the strength of current techniques to more accurately identify recent galaxy formation (e.g., color-magnitude diagrams from photometric surveys which lead to estimates for age and star formation histories and phase-space diagrams from low-res spectroscopic surveys which lead to estimates for accretion histories), it is encouraging that our technique works so well for early accretion epochs and low luminosity objects.

In the last column of Figure 7, we present a summation of estimates across accretion epochs (shown with  $\langle \text{FoE} \rangle$  values labeled “L”) and across binned satellite luminosities (labeled “T”) for all epochs. Here, we confirm that a marginalization of estimates across our two epochs yields 1-D estimates with greater fidelity than its 2-D decomposition for the *worst* EM estimates as indicated by the L-labeled  $\langle \text{FoE} \rangle$  values. More importantly, we can compare our best *worst* values for our h7 estimates ( $\text{FoE} = 2.161$ ) to the respective 1-D h8 estimates ( $\text{FoE} = 2.059$ ) in Figure 4. A comparison of these values shows tentative evidence that our hypothesis about gains in STS information is correct — that the 1-D marginalizations across epochs from a 2-D STS provides on par or better estimates for 1-D AHP than does our *bona fide* 1-D STS. We can also compare the set of “T”-labelled best  $\langle \text{FoE} \rangle$  values for our 1-D marginalizations across satellite luminosity bins in Figure 7 to the set of values calculated for Figure 5 ( $\text{FoE} = [7.168, 485.6]$  for our *best* and *worst* values, respectively). Here, we find that our estimates for our time of accretion histories improve substantially overall, and dramatically when comparing our *best* and *worst* AHP estimates. The next two sections address whether these improvements are ubiquitous as we increase the resolution of our STS in the accretion time dimension.

#### 4.1.2. “Medieval” accretion: 3x5 STS results



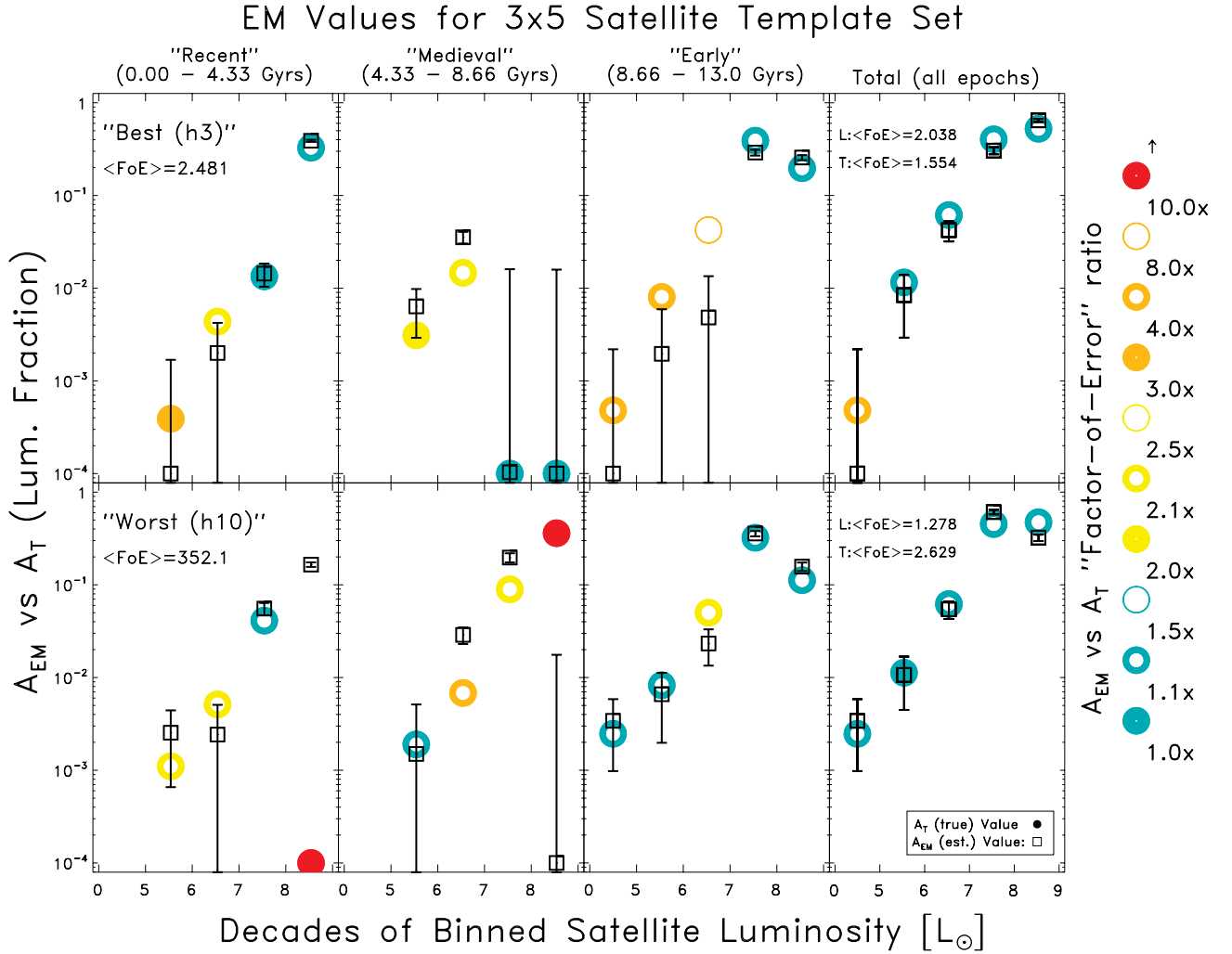
**Figure 7.** Figure of 2x5 STS is similar to Figure 4 but first two columns shows separate sets of templates for *recent* (0–6.5 Gyrs) and *early* (6.5–13.0 Gyrs) accretion epochs. Final column shows totals over all time (i.e., an “effective” 1x5 STS from adding corresponding estimates from both columns). Numbers labeled “L” and “T” refer to  $\langle \text{FoE} \rangle$  values calculated across satellite stellar mass and time bins, respectively.

In order to further test our ability to estimate AHPs, we seek to increase our accretion time resolution (by adding an intermediate “medieval” accretion epoch), with the hopes that greater information from an expanded STS will lead to better AHP estimates.

Figure 8 shows our *best* and *worst* 3x5 STS results. The  $\langle \text{FoE} \rangle$  values between the *best* and *worst* EM estimates show a substantial decrease in quality. It is immediately apparent (from color) that individual estimations fared significantly worse than they did in the 2x5 STS selections of Figure 7. Also, by inspection, the *medieval* epoch yields the worst estimates overall. Similar to Figure 7, early epoch estimates of Figure 8 are the most accurate. The overall decrease in performance from our 2x5 to 3x5 STS is likely due to the degeneracy in CARD space between some adjacent templates in the 3x5 STS (e.g., see Figure 3 for illustration of this effect) and across accretion time for the higher luminosity templates. For example, if we look across the *recent* and *medieval* epochs for our *worst* EM estimate selection, we can see that there are degeneracies in the estimates for the highest stellar mass bins ( $10^8$ – $10^9 M_\odot$ ). These degeneracies are due to the increasing similarities between

chemical model tracks of more massive (and luminous) dwarf satellite models. Such degeneracies can lead to the satisfaction of estimates across all epochs by one individual template (e.g., h7 from Fig. 5), by distributing the luminosity fraction amongst co-degenerate templates (e.g., h7 from Fig. 7), or by swapping estimates across adjacent epochs (e.g., h10 from Fig. 8). However, it appears that a clear separation in accretion epochs for the same stellar mass bins possibly reduces degeneracies between them (as seen for the *best* (h3) estimates).

If we look at the final column for our 1-D marginalizations from the 2-D 3x5 STS, we once again see improvements in  $\langle \text{FoE} \rangle$  values in comparison to Figures 4 and 5 (e.g., look at “L” and “T” values for all selections versus uniformly-weighted values in Fig. 12 of §5). While improvements were anticipated, it is still surprising, given the relative lack of success for individual 3x5 STS templates, that marginalization of the *worst* 3x5 STS leads to 1-D estimates that offer an improvement over the 2x5 STS marginalized 1-D estimates. In this case, some inaccuracies due to degeneracies across epochs are mitigated by summation over accretion epochs. Consequently, improvements to our marginalized *mass-resolved* 1-D es-



**Figure 8.** Figure of 3x5 STS is similar to Figure 7 but includes an additional column for an intermediate *medieval* accretion epoch.

estimates arise from an increase in the STS epoch resolution. Presumably, the better estimates would originate directly from improved individual epoch estimates. However, poor individual estimates due to degeneracies within the same stellar mass bins refute this idea. Indeed, it is more likely that improvements to our epoch resolution led to better estimates indirectly, by not decreasing the degeneracies between adjacent epochs, but rather decreasing degeneracies between adjacent stellar mass bins. While the effects described above are certainly taking place, it is still unclear from Figures 4, 5, 7, and 8 whether these improvements remain across all 11 halos. In the next section we examine the  $\langle \text{FoE} \rangle$  values as ensembles across the 11 halos to determine the overall success of recovering AHPs given our STS.

#### 4.2. Comparison of results across all STS

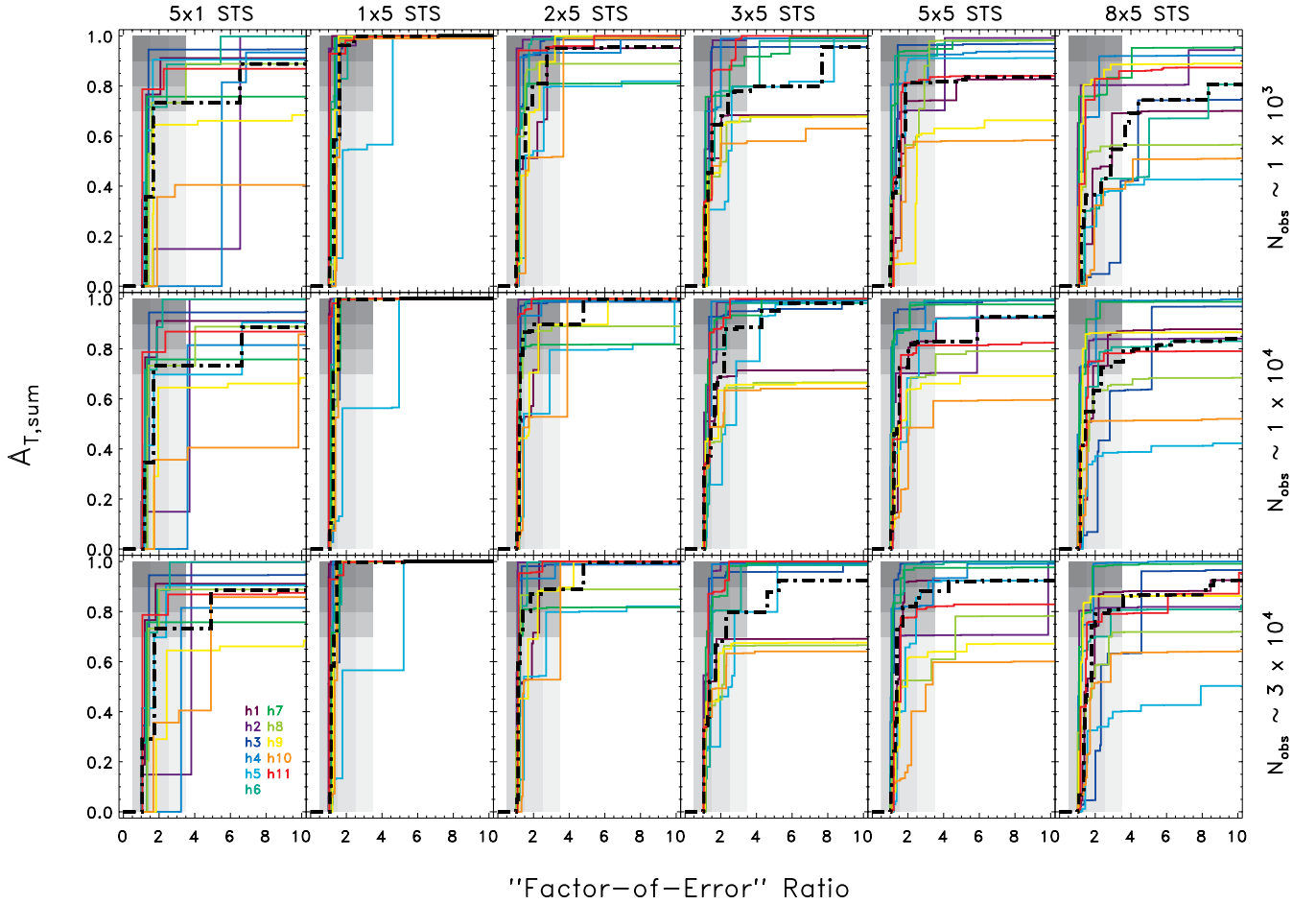
In this section we compare results from all our simulated halos and the templates we constructed. Using FoE values (see §2.5) we can determine a cumulative distribution function (CDF) of FoE values with respect to  $A_T$  for each STS used. The CDF values described above (which we call  $A_{T,sum}$ ) indicate the fraction of the total stellar halo mass we can identify within a given FoE value.

First, we construct  $A_{T,sum}$  values in Figure 9 for six of our 10 STS. Each plot frames the recovery of AHPs in

terms of the level of accuracy (i.e., FoE) at which we can characterize a certain portion ( $A_{T,sum}$ ) of the total luminous stellar content of the halos examined. Once again differences in the fidelity of our estimates between 5x1 and 1x5 STS are clearly shown with a median  $A_{T,sum}$  (fraction recovered) with a FoE  $\lesssim 2$  being  $\simeq 73\%$  and  $95\text{--}99\%$ , respectively. Characterizing the success of the method overall, we find that the median  $A_{T,sum}$  (with FoE  $\lesssim 2$ ) across most STS is  $\simeq 70\%$  or better. It is evident from the STS shown in Figure 9 that EM estimates fair poorly when applied to certain halo realizations. We discuss possible causes for the often poorer estimates of a few halos in §5.

Figure 10 displays another way we can summarize our results with the utilization of  $A_{T,sum}$  and FoE. In the three panels, box-and-whisker plots illustrate the median and shape of the distribution of  $A_{T,sum}$  values calculated for estimates with FoE  $\lesssim 2$  amongst all 11 halos.<sup>6</sup> The top panel displays similar information to the results

<sup>6</sup> The actual chosen cutoff here for FoE values is  $\leq 2.25$ . Given that this research is presented as a proof-of-concept, we wanted to capture FoE values that were consistent with a FoE = 2. Since such a cutoff is arbitrary, the reader is free to reexamine the selected columns of Figure 9 and reconstruct  $A_{T,sum}$  estimates for different FoE cutoff values.



**Figure 9.** Figure displays six STS-derived plots of  $A_{T,sum}(\leq \text{FoE})$  for all 11 halos demonstrating another benchmark for our CARD analysis for deriving the AHPs of our halos. Columns represent results for listed STS estimates. Rows represent estimates derived from a certain number of observed stars which are labeled at the right edge of each row. Shaded areas in each plot guide the eye to FoE estimates of  $\sim 2 - 3$  or better which primarily indicate estimates that cover  $A_{T,sum} \gtrsim 70\%$ . Individual solid colored lines represent each of the 11 halos used in the study. Colored labels for the halos are shown in the bottom-left of the figure. Black dot-dashed CDF represent the median of all 11 halos vs. FoE values.

shown in Figure 9. The middle and bottom panels show both genuine and marginalized estimates for the 1x5 STS accreted mass functions and the 5x1 STS accretion time histories, respectively.

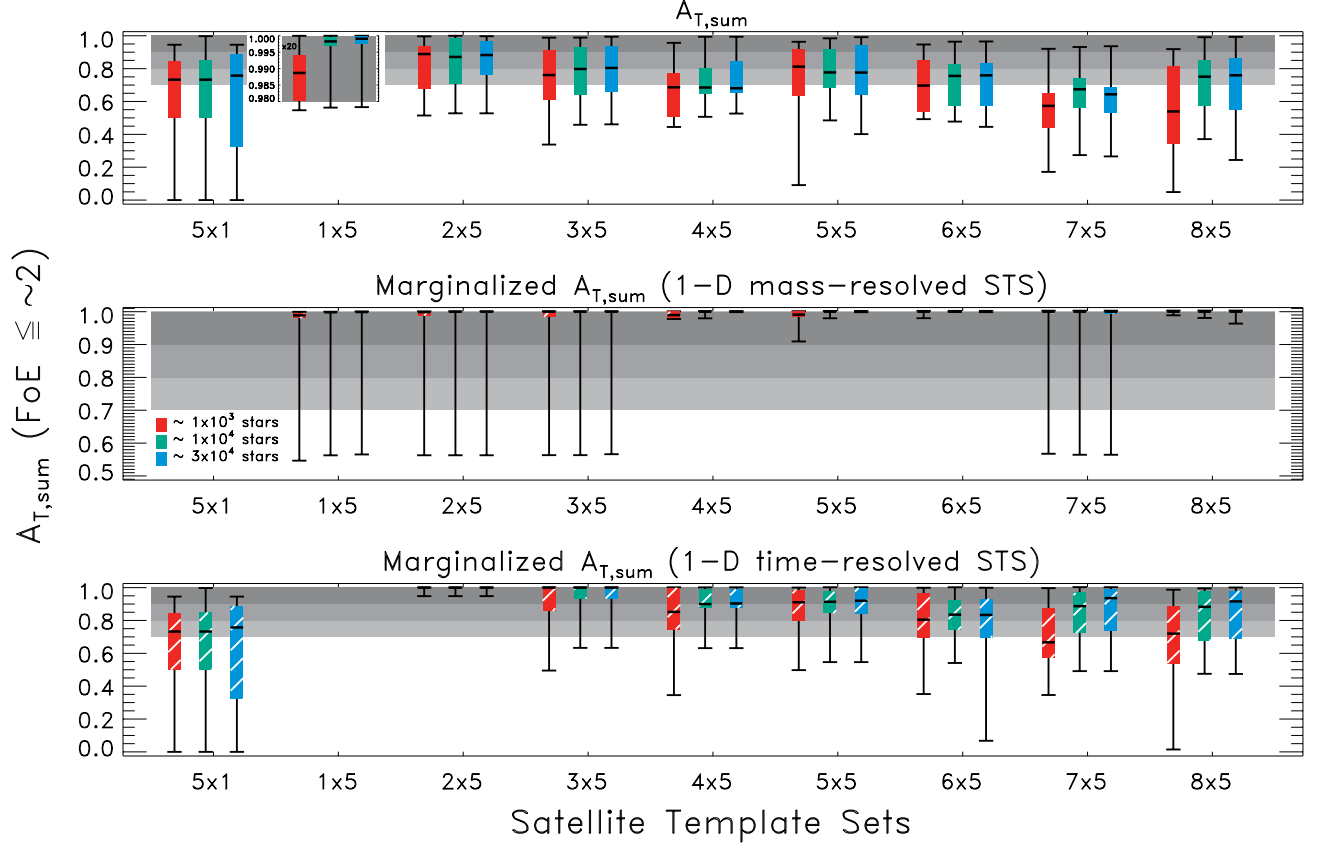
In the top panel,  $A_{T,sum}(\text{FoE} \lesssim 2)$  is plotted, as a color box, for all STS examined. Here, as in Figure 6, the color refers to the respective number of observations used (as indicated in the plot legend). In the plot, we see that our best median  $A_{T,sum}$  values are given by the 1x5 and 2x5 STS while the worse values are given by 5x1 and 7x5 STS. The average among the best and worst  $A_{T,sum}$  values across all STS examined and for increasing number of stellar observations are  $\sim 0.96 - 0.98$  and  $\sim 0.29 - 0.41$ , respectively. The average median  $A_{T,sum}$  values across all STS examined and for increasing number of stellar observations are 0.742, 0.783, and 0.785. This means that on average our FoE are  $\lesssim 2$  for at least  $\sim 75\%$  of the total halo stellar mass (i.e.  $A_{T,max} = 0.75$ ) observed.

Marginalized values, which are defined in §4.1.1, are useful for evaluating any gains that may potentially arise due to better time (or mass) resolution. More precisely, any information about templates that is lost or gained should generally result in a corresponding rise or drop in  $\langle \text{FoE} \rangle$  and thus appear as an increase in  $A_{T,sum}(\text{FoE} \lesssim 2)$ . As a reference, a grey bar is placed in each panel

to indicate a region where the  $A_{T,sum}(\text{FoE} \lesssim 2)$  values range from 70% to 100% (from bottom to top).

The middle panel shows our *mass-resolved* marginalized values (summed over accretion time bins) for 8 of the 9 STS (with 5x1 omitted because its value is not applicable in this context). The plot shows an across-the-board increase in  $A_{T,sum}(\text{FoE} \lesssim 2)$  values (i.e., a general drop in all STS  $\langle \text{FoE} \rangle$  values) measured for a recovery of the total stellar mass function. The improvement in  $\langle \text{FoE} \rangle$  despite the tendency for individual FoE STS values to increase with an increase in the number of templates used indicates that significant gains were made by using a larger template set for the specific purpose of generating more accurate estimates of a halo's total stellar mass function (via marginalization).

The bottom panel shows our *time-resolved* marginalized values (summed over mass bins) for 8 of the 9 STS (with 1x5 also omitted because its value is not applicable in this context). In this case, the plot shows a descending trend in  $A_{T,sum}(\text{FoE} \lesssim 2)$  values with larger STS (i.e., a generally ascending rise in  $\langle \text{FoE} \rangle$  values with increasing STS size) measured for a recovery of the total accretion time history. Despite the decrease  $A_{T,sum}(\text{FoE} \lesssim 2)$  values, these values remain relatively good (above 70% for  $A_{T,sum}$  values above the bottom 50% margin) up to our



**Figure 10.** Figure shows box-and-whisker plots of  $A_{T,sum}(\text{FoE} \lesssim 2)$  for full STS (top), marginalized 1-D *mass-resolved* STS (middle), and marginalized *time-resolved* (bottom) using all STS examined for our 11 halos. The median values of  $A_{T,sum}$  for all 11 halos are shown as a black line across every box. The 25<sup>th</sup> and 75<sup>th</sup> percentiles of the distribution are shown as the lower and upper bounds of the each box, respectively. Whiskers designate the minimum and maximum values for  $A_{T,sum}$  values in the distributions shown. Each box has a color that refers to the number of stars identical to the colors used in Figure 6. *Top:* Boxes in the top panel (solid colors) refer to the genuine  $A_{T,sum}$  values for each respective STS. *Middle and Bottom:* “Marginalized” boxes (striped colors) refer to the  $A_{T,sum}$  values calculated from the sum across the mass (time) dimension of templates into an effective 1x5 (5x1) template (e.g., see Figures 7 and 8). 1x5 STS (*mass-resolved*)  $A_{T,sum}$  values derived from marginalizing over time-binned estimates are shown in the middle panel while 5x1 STS (*time-resolved*)  $A_{T,sum}$  values derived from marginalizing over mass-binned estimates are shown in the bottom panel. Increasingly darker grey bands spanning all STS (for  $70\% \leq A_{T,sum} \leq 100\%$ ) are shown to highlight the success of our estimates.

6x5 STS. Indeed, all *time-resolved* marginalized values show a significant improvement in accretion time histories over the history given by the 5x1 STS. Overall, the results show that we could expect to recover accretion time histories using the EM algorithm given that we use reasonable templates.

Results shown in Figure 9 and Figure 10 prove that even with the simplest template divisions, we could, with the appropriate dataset, recover the accretion history of the MW halo. To that point, we find that these STS EM estimates can recover the total contributions from accreted systems (templates) of similar mass (i.e. halo luminosity function) to within a factor of 1.02 ( $\leq 2\%$  of the *true* value) for most of the 11 halos. Separately, the EM algorithm can determine the mass fractions within accretion times to within a factor of  $\gtrsim 4$  for at least 90% of the halo’s total stellar mass. Both results present encouraging prospects for recovering the accretion history of the MW halo from current and near-future data collections.

## 5. DISCUSSION

In the following discussion, we examine the statistical reliability (or robustness) of the EM algorithm when ap-

plied to our models and simulated data. We also explore what masses the current approach is most sensitive to and discuss implications for future work.

### 5.1. Reliability

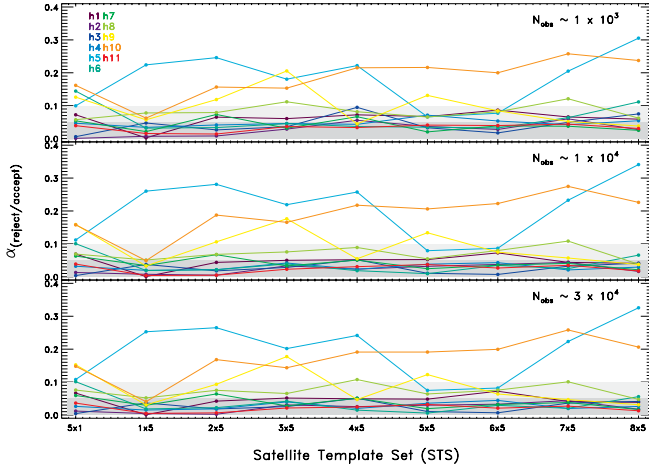
We can test the statistical robustness of the EM algorithm’s application to our simulated halos by performing a *likelihood ratio test* on the results of our analysis. By determining the true ( $\mathbf{A}_T$ ) and respective  $\mathbf{A}_{EM}$  likelihood values from each application of STS to our halos via the EM algorithm, we can calculate a  $\chi^2$ -statistic defined by the following equation

$$\chi^2 = -2 \ln\left(\frac{\lambda_T}{\lambda_{EM}}\right) \quad (7)$$

where  $\lambda_T$  and  $\lambda_{EM}$  are the likelihoods for  $\mathbf{A}_T$  and  $\mathbf{A}_{EM}$  values, respectively. One can then reject the assumption that the true AHP templates are well-approximated by the STS used if the  $\chi^2$ -value from Eqn. 7 is larger than the  $\chi^2$ -percentile values given  $k$  degrees-of-freedom ( $k = m_{EM} - m_T$ )<sup>7</sup> and a confidence level denoted by  $\alpha$ . Fig-

<sup>7</sup> Hence  $k$  equals the number of templates in a STS estimate ( $m_{EM}$ ) minus the number of those templates that are actually





**Figure 11.** Figure shows the  $\alpha$ -level threshold for accepting or rejecting the null hypothesis that suitable AHP templates were used in estimating  $\mathbf{A}_T$  values. Colors represent results for the 11 halos examined and panels compare results for the approximate number of stars observed. See text for discussion.

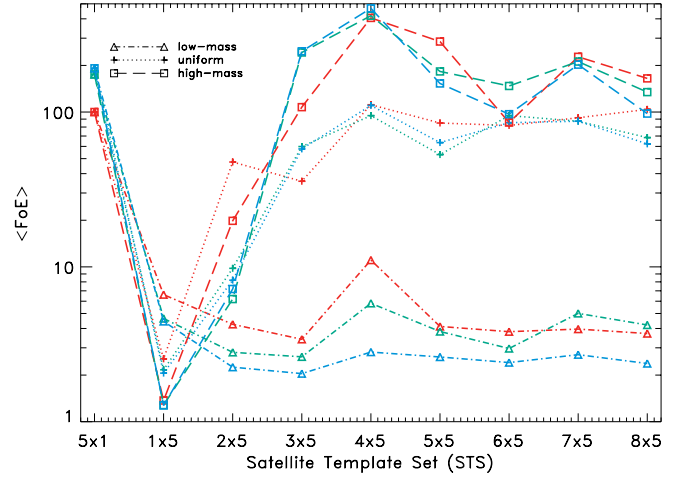
ure 11 shows the maximum  $\alpha$ -value one can assume for a  $\chi^2$ -distribution before you have to reject the assumption that suitable AHP templates are chosen. For example, an  $\alpha = 0.05$  corresponds to a confidence that 95% of all samples taken of a given size are well characterized by the STS in use. Here, we find that all sample sizes and STS used, halos 5, 9, and 10 are by far the worst characterized halos by our STS divisions. For most STS used, these halos are ill-matched to the generic STS created in our division scheme and therefore challenge the robustness of this method. Such challenges need to be address before this method can be utilized to model the AHP of the MW halo. The solution resides in the development and incorporation of sufficiently realistic models of dwarf CARdS into this method — a goal that will be addressed in future work.

### 5.2. Sensitivity to different mass bins

Another consideration in assessing the reliability of our method is to determine how well it uncovers AHPs based on the satellite mass regime we are interested in. Taking Eqn. 6 from §2.5, we can calculate  $\langle \text{FoE} \rangle$  values with different weights — i.e., uniform (mean), low-mass preferred, or high-mass preferred — based on what satellite population(s) one prefers to recover. Figure 12 shows the median  $\langle \text{FoE} \rangle$  amongst all halos for each STS used. The same colors from Figure 10 are used indicate the number of stars used for the analysis and symbols and corresponding lines refer to the type of weighting used (see figure legend). Uniformly-weighted  $\langle \text{FoE} \rangle$  values are weighted by  $m^{-1}$  (i.e. by the number of templates used) and identical to the weighting used for the main results of this paper. Weights that emphasize more accuracy in low- or high-mass satellite AHPs are weighted by the corresponding upper bin mass limits and their reciprocals, respectively.

In the figure, we can see that  $\langle \text{FoE} \rangle$  values for low-mass satellite recovery fair the best whereas uniform and high-mass satellite recovery-emphasized weights are a factor of  $\gtrsim 10$  in all but the three smallest template sets. In other words, when one emphasizes the accurate recovery of

occupied in the true AHP ( $m_T$ ).



**Figure 12.** Figure shows  $\langle \text{FoE} \rangle$  values for different template weights. The various colors refer to the approximate number of stars used as indicated in Fig. 10. Weights are listed in the figure legend. See text for discussion.

low-mass satellites, the weighting favors templates with lower FoE values which yields lower overall  $\langle \text{FoE} \rangle$  values. This result further clarifies the immediate strengths of the method: its adept at differentiating between accreted dwarfs of low-mass in CARD-space due to the lack of degeneracies in their occupied region of space. Meanwhile, its clear that while degeneracies exist in the CARD-space occupied by high-mass satellites and larger STS, we are encouraged by the fact that the introduction of more templates can significantly decrease degeneracies in only two dimensions of CARD-space.

### 5.3. Future Prospects

It is clear from both our results and our reliability tests that the current method fails often for three of the 11 halo simulations. From our examination of these three problematic halos we find that all of them show predominately early accretion of massive dwarf galaxies with integrated CARdS that appear to be highly-degenerate when compared to the other eight halos AHP CARdS examined. To address the degeneracies that exist (particularly among high-mass systems) we posit that differences between mass-dependent (nucleosynthetic) yields for different nucleosynthetic sites and elements groups (e.g., see Lee et al. 2013) can be exploited to greatly reduce or remove such degeneracies by expanding the CARD-space basis set.

For example, we only looked at two dimensions in CARD space whereas more recent work on “chemical tagging” expands the number of dimensions available by establishing the best chemical abundance signatures to pursue in chemical abundance space in order to optimize survey efforts (e.g., the GALAH survey). One way to optimize our surveys for searches in chemical abundance space is to prioritize spectroscopic observations for elements that confer the greatest amount of distinction between systems with different origins. To this end principle component analysis (PCA) was used by Ting et al. (2012) to identity and rank the 6 – 9 most distinguishing elements in chemical abundance space. In their work, the chemical abundance space of various parts of both the galaxy and the galactic neighborhood were examined to determine the best elements to observe in order to de-

cipher their galactic chemical evolution. A CARD-space basis set derived from various combinations of these elements are likely to offer the breaks in degeneracies that we require.

## 6. SUMMARY

In our investigation to determine the efficacy of recovering the accretion history of the MW halo, we used simulated halo data from the Bullock & Johnston (2005) MW halo simulations. Our approach required the CARDS of  $[\alpha/\text{Fe}]$  and  $[\text{Fe}/\text{H}]$  for the 11 simulated realizations for accretion-grown halos, observed samples of stars from those simulations, and CARD templates of accreted dwarfs models in the simulations. From this assortment of data we were able to apply a statistical algorithm (the EM algorithm) which utilizes the model templates with those *observed* stars to disentangle the accretion history of our simulated halos.

To evaluate the success of our estimates, we examined relationships between a measure of accuracy, the FoE, and a measure of the maximum fraction of the halo’s stellar mass that is characterized by this level of accuracy which we call  $A_{T,sum}$ .

In our analysis, we employed (equally-partitioned) STS as model sets for our generative mixtures (i.e., the simulated halos). The first test of our templates involved 1-D STS which were composed entirely of either stellar mass or accretion time partitions. In the case of our 1-D *mass-resolved* STS, the EM algorithm estimates for individual templates were made to within a factor of 8 (in the worst case) for halo 5 and were within a factor 1.5 – 2.5 or better for most mass bins. However, in the case of our 1-D *time-resolved* STS, results were considerably less accurate, with approximately half of the individual templates being off by a factor of 10 or more. In this case, it is important to note that the bulk of these poor estimates occurred for bins containing the least amount of accreted mass. This outcome was not unexpected, but it stands in sharp contrast to estimates that resulted from our *mass-resolved* case. In both cases, we also examined the effect of increasing our datasets from one thousand to thirty thousand stellar chemical abundance observations. While we found that an increase in our data generally led to better estimates from our *mass-resolved* templates no improvement was seen for estimates from our *time-resolved* templates. These results lead us to examine what, if any, improvements could be made in our EM estimates by expanding our STS into two dimensions of accretion time and mass and increasing the number of templates used.

In examining the use of the 2-D STS in EM algorithm estimations, we find that these template sets provided more accurate estimates in general. More precisely, we find that our 2x5 STS could be used to furnish remarkably good AHP estimates — meaning that we could easily recover a tally of satellites that fell in *recently* versus those that fell in more than 6.5 Gyrs ago. It is clear that in this dichotomous evaluation mode, the EM algorithm can easily detect distinction between previous satellites that were accreted from 6.5 Gyrs ago to now and those satellites that accreted prior to that time using only two dimensions in chemical abundance space. Also, we find that in the case where we try to estimate an *early*, *medieval*, and *recent* accretion history — our

3x5 STS tests — the EM estimates do fairly well too. In some cases it was apparent from our 2-D STS figures (for our 3x5 STS in particular) that degeneracies between templates in a set were possibly degrading our EM estimates and perhaps limiting the potential for this technique. However, despite such degeneracies, we find that we can improve our 1-D recovery of both the mass accretion history (functionally similar to mass/luminosity functions) and the accretion time history (a coarse account of mass growth of the halo over time) by marginalizing estimates across templates in the appropriately related dimension. Thus, we are confident that at the very least this technique can be used, albeit carefully, to produce fairly accurate estimates for 1-D accretion mass or mass growth functions for the MW halo.

Finally, we compare our tests for all 2-D STS. We find three interesting features that reflect the technique’s potential. These features are: (1) fairly accurate estimates for AHPs across most STS used (2) consistent or improved 1-D *mass-resolved*  $A_{T,sum}$  values from 1-D marginalization over an increase in the number of templates used, and (3) a substantial overall improvement in the marginalized *time-resolved*  $A_{T,sum}$  values across all STS used over the 1-D 5x1 STS values. From these features we conclude that, on average, we can recover the bulk of accreted dwarfs’ relative contributions to the halo’s accretion history by mass, to within a factor of  $\sim 2$ . Despite this fact, many individual templates (especially our lower mass bin templates) can produce estimates that are far less accurate than estimates given for the main stellar mass contributors to the halo. This is likely due to degeneracies among templates belonging to same STS and relative contributions of these objects to the general star count of halo. These issues that can be addressed by carefully selecting which observed stars are to be included in the data sample and by expanding the chemical abundance space basis set to better disentangle the individual star formation histories of the previously accreted dwarf satellites in our halos (or our Halo).

Lastly, in spite of the demonstrated drawbacks involving degeneracies between individual templates, we find that, remarkably, it is possible to improve 1-D mass function predictions (as a function of accreted satellite mass or accretion time) simply by increasing the number of partitioned time bins (templates) used for EM estimates and then marginalizing over those estimates in either stated dimension. This result means that at the very least it is possible to extract, e.g., accurate luminosity functions with estimates that clearly improve with better resolution in our  $t_{acc} - M_{sat}$  plane. Further investigation of this result will be pursued in the near future.

## 7. CONCLUSIONS

In conclusion we note the following implications of our study:

- Our proof-of-concept is verified — recovering halo accretion histories using their CARD information works (and works well for a certain level of detail)
- In particular, even when applying our method to only 2-D CARD-space we appear to be sensitive to:
  - early accretion events (regions where information in phase-space has phase-mixed away)

- low luminosity dwarfs (objects we cannot see *in-situ* because they are too faint)

- There *are* degeneracies in 2-D CARD-space, particularly amongst high mass accreted dwarfs
- However, since we only looked in 2-D and there are prospects of 10's of thousands of stars with  $> 6$  independent chemical dimensions it is very important to pursue this method of approach further

Finally, given these implications we are compelled to generate more realistic templates from chemical evolution models in higher dimensions and test them against existing dwarf data. It is the hope that by validating the fidelity of such templates, we could, in turn, employ these templates in our method to produce a detailed account of the accretion history of the MW halo.

DML would like to give thanks to his dissertation thesis committee for their helpful comments and support in the writing of this paper. KVJ and DML would also like to give thanks to James Bullock, Brant Robertson and Andreea Font for the collaboration that developed the numerical data sets used in this work. This work was supported by the “973 Program” 2014 CB845702; the Strategic Priority Research Program “The Emergence of Cosmological Structures” of the Chinese Academy of Sciences (CAS; grant XDB09010100). DML was also supported by NSF grants AST-0806558 and AST-1107373.

## REFERENCES

- Belokurov, V. et al. 2006, ApJ, 642, L137
- Bland-Hawthorn, J., & Freeman, K. C. 2004, Publications of the Astronomical Society of Australia, 21, 110
- Bland-Hawthorn, J., Karlsson, T., Sharma, S., Krumholz, M., & Silk, J. 2010, ApJ, 721, 582
- Bonifacio, P., Sbordone, L., Marconi, G., Pasquini, L., & Hill, V. 2004, A&A, 414, 503
- Bullock, J. S., & Johnston, K. V. 2005, ApJ, 635, 931
- Cayrel, R. et al. 2004, A&A, 416, 1117
- De Silva, G. M., Freeman, K. C., Asplund, M., Bland-Hawthorn, J., Bessell, M. S., & Collet, R. 2007, AJ, 133, 1161
- Efstathiou, G., Davis, M., White, S. D. M., & Frenk, C. S. 1985, ApJS, 57, 241
- Eggen, O. J., Lynden-Bell, D., & Sandage, A. R. 1962, ApJ, 136, 748
- Font, A. S., Johnston, K. V., Bullock, J. S., & Robertson, B. E. 2006, ApJ, 638, 585
- Freeman, K., & Bland-Hawthorn, J. 2002, ARA&A, 40, 487
- Fulbright, J. P. 2002, AJ, 123, 404
- Geisler, D., Smith, V. V., Wallerstein, G., Gonzalez, G., & Charbonnel, C. 2005, AJ, 129, 1428
- Geisler, D., Wallerstein, G., Smith, V. V., & Casetti-Dinescu, D. I. 2007, PASP, 119, 939
- Helmi, A., & de Zeeuw, P. T. 2000, MNRAS, 319, 657
- Ibata, R. A., Gilmore, G., & Irwin, M. J. 1994, Nature, 370, 194
- Ivans, I. I., Sneden, C., Kraft, R. P., Suntzeff, N. B., Smith, V. V., Langer, G. E., & Fulbright, J. P. 1999, AJ, 118, 1273
- Johnson, J. A., Ivans, I. I., & Stetson, P. B. 2006, ApJ, 640, 801
- Jonsell, K., Edvardsson, B., Gustafsson, B., Magain, P., Nissen, P. E., & Asplund, M. 2005, A&A, 440, 321
- Kaufer, A., Venn, K. A., Tolstoy, E., Pintte, C., & Kudritzki, R.-P. 2004, AJ, 127, 2723
- Lacey, C., & Cole, S. 1993, MNRAS, 262, 627
- Lee, D. M., Johnston, K. V., Tumlinson, J., Sen, B., & Simon, J. D. 2013, ApJ, 774, 103
- Majewski, S. R. et al. 2005, AJ, 130, 2677
- Majewski, S. R., Skrutskie, M. F., Weinberg, M. D., & Ostheimer, J. C. 2003, ApJ, 599, 1082
- Monaco, L., Bellazzini, M., Bonifacio, P., Ferraro, F. R., Marconi, G., Pancino, E., Sbordone, L., & Zaggia, S. 2005, A&A, 441, 141
- Newberg, H. J. et al. 2002, ApJ, 569, 245
- Nissen, P. E., & Schuster, W. J. 1997, A&A, 326, 751
- Pompeia, L. et al. 2006, ArXiv Astrophysics e-prints
- R. G. Gratton, E. Carretta, R. Claudi, S. Lucatello, & M. Barbieri. 2003, A&A, 404, 187
- Robertson, B., Bullock, J. S., Font, A. S., Johnston, K. V., & Hernquist, L. 2005, ApJ, 632, 872
- Schlaufman, K. C. et al. 2009, ApJ, 703, 2177
- Schlaufman, K. C., Rockosi, C. M., Lee, Y. S., Beers, T. C., Prieto, C. A., Rashkov, V., Madau, P., & Bizyaev, D. 2012, Astrophysical Journal, 749, 77
- Searle, L., & Zinn, R. 1978, ApJ, 225, 357
- Sharma, S., Johnston, K. V., Majewski, S. R., Muñoz, R. R., Carlberg, J. K., & Bullock, J. 2010, ApJ, 722, 750
- Shetrone, M., Venn, K. A., Tolstoy, E., Primas, F., Hill, V., & Kaufer, A. 2003, AJ, 125, 684
- Shetrone, M. D., Côté, P., & Sargent, W. L. W. 2001, ApJ, 548, 592
- Smecker-Hane, T. A., & McWilliam, A. 2002, ArXiv Astrophysics e-prints
- Somerville, R. S., & Kolatt, T. S. 1999, MNRAS, 305, 1
- Stephens, A., & Boesgaard, A. M. 2002, AJ, 123, 1647
- Tautvaisienė, G., Geisler, D., Wallerstein, G., Borissova, J., Bizyaev, D., Pagel, B. E. J., Charbonnel, C., & Smith, V. 2007, AJ, 134, 2318
- Ting, Y.-S., Freeman, K. C., Kobayashi, C., De Silva, G. M., & Bland-Hawthorn, J. 2012, MNRAS, 421, 1231
- Unavane, M., Wyse, R. F. G., & Gilmore, G. 1996, MNRAS, 278, 727
- Venn, K. A. et al. 2001, ApJ, 547, 765
- Venn, K. A., Tolstoy, E., Kaufer, A., Skillman, E. D., Clarkson, S. M., Smartt, S. J., Lennon, D. J., & Kudritzki, R. P. 2003, AJ, 126, 1326
- York, D. G. et al. 2000, AJ, 120, 1579

## APPENDIX

### THE EXPECTATION-MAXIMIZATION ALGORITHM

#### *Expectation step*

To implement the algorithm, we first need to derive the expression for the complete data log likelihood, given by Eqn. 4, which is conditioned on the data. To do this, it is necessary to decide on a mode of usage for  $z_{ij}$ . The use of  $z$  casts the EM algorithm as either *hard* when its value discretely indicates the  $f_j(x_i, y_i)$  of origin or *soft* when its value probabilistically indicate the origin of point  $(x_i, y_i)$  across all  $f_j$ . For this application, we chose to implement a *hard* EM algorithm for estimation of  $A_{MLE}$  in which  $z_{ij}$  has an true value equal to **1** if the data point  $(x_i, y_i)$  comes from

model  $f_j$  or  $\mathbf{0}$ , otherwise. Thus our overall expectation is

$$E_{\mathbf{A}}[\ell(\mathbf{A})|\mathbf{x}, \mathbf{y}] = \sum_{i=1}^n \sum_{j=1}^m E_{\mathbf{A}}[z_{ij}|x_i, y_i] \{\log A_j + \log f_j(x_i, y_i)\} \quad (\text{A1})$$

where

$$E_{\mathbf{A}}[z_{ij}|x_i, y_i] = \frac{A_j f_j(x_i, y_i)}{\sum_{k=1}^m A_k f_k(x_i, y_i)} \quad (\text{A2})$$

as defined by Eqn. 2. Since we are ultimately maximizing Eqn. A1, the non-constant term, Eqn. A2, becomes the component of interest. To iteratively evaluate this expectation, we let  $w_{ij}^{(t)}$  be Eqn. A2 at the  $t^{\text{th}}$  step:

$$w_{ij}^{(t+1)} = \begin{cases} \frac{A_j f_j(x_i, y_i)}{\sum_{k=1}^m A_k f_k(x_i, y_i)} & j = 1, \dots, m \\ 1 - w_{i1} - \dots - w_{i,m-1} & j = m. \end{cases}$$

Since  $\mathbf{A}$  is not defined for the first evaluation, we use a random initialization to generate  $\mathbf{w}_j^{(0)}$ . Here, it should be noted that convergence is not sensitive to the choice of values in our case, though it can be in cases where the likelihood is riddled with local maxima. If we examine the expression above, we can conceptually define the mechanism for maximization as a “ratcheting up” of  $E_{\mathbf{A}}[z_{ij}|x_i, y_i]$  values by maximizing  $A_j f_j(x_i, y_i)$  with respect to  $\sum_{k=1}^m A_k f_k(x_i, y_i)$ . Derivation of the maximization expression is discussed below.

#### Maximization step

Above we defined an explicit formulation for the expected log-likelihood (Eqn. A2) given a single parameter  $\mathbf{A}$  and the data  $(\mathbf{x}, \mathbf{y})$ . The argument of the maximum of Eqn. A2 at each iteration  $t$  provides an estimate that approaches the MLE of  $\mathbf{A}$ , and is given by:

$$\mathbf{A}^{(t)} = \underset{\mathbf{A}}{\operatorname{argmax}} [\ell(\mathbf{A})|\mathbf{x}, \mathbf{y}, \mathbf{A}^{(t-1)}]. \quad (\text{A3})$$

Accounting for the  $m-1$  free parameters of  $\mathbf{A}$ , differentiation of Eqn. A1 with Eqn. A2 proceeds, for  $k = 1, \dots, m-1$ , as:

$$\frac{\partial}{\partial A_k} E_{\mathbf{A}}[\ell(\mathbf{A})|\mathbf{x}, \mathbf{y}] = \sum_{i=1}^n \left\{ w_{ik}^{(t-1)} \frac{1}{A_k} - w_{im}^{(t-1)} \frac{1}{1 - A_1 - \dots - A_{m-1}} \right\}$$

where the first term in the summation accounts all values of  $k \leq m$  and the second term eliminates over-counting of the  $1^{\text{st}}$ -term at  $k = m$ . The derivative of an argmax is always equal to zero since we are taking a derivative at the maximum point of the function in question (in our case the expectation of the log-likelihood). Thus, we can expand the summation of data points and equate the terms described above to one another

$$\frac{1}{A_k} \sum_{i=1}^n w_{ik}^{(t-1)} = \frac{1}{1 - A_1 - \dots - A_{m-1}} \sum_{i=1}^n w_{im}^{(t-1)}.$$

Consequently, these terms being equal means that every  $k \leq m$  term is equal to each as shown below

$$\frac{1}{A_k} \sum_{i=1}^n w_{ik}^{(t-1)} = \dots = \frac{1}{A_{m-1}} \sum_{i=1}^n w_{i,m-1}^{(t-1)} = c$$

and

$$\mathbf{A}_k^{(t)} = \frac{\sum_{i=1}^n w_{ik}^{(t-1)}}{c}$$

where  $c$  is some constant.

The unknown constant  $c$  appears problematic, but, because  $\sum_{j=1}^m A_j = 1$ , algebraic manipulation reveals that  $c = n$ , yielding a final solution that can be numerically evaluated:

$$\mathbf{A}_k^{(t)} = \frac{\sum_{i=1}^n w_{ik}^{(t-1)}}{n} \quad (\text{A4})$$

$$\mathbf{A}_m^{(t)} = 1 - A_1 - \dots - A_{m-1}. \quad (\text{A5})$$

Finally, to implement this algorithm, we simply compute an initial value for  $\mathbf{A}$ , inserting each component,  $A_j$ , into an  $w_{ik}^t$  equal to Eqn. A2 (i.e. with  $k$  initially identical to  $j$ ) and then compute that expression with Eqn. A4 to calculate each new corresponding  $A_k$ . This process is repeated unto our iteration criterion is met.

In our case, computation of  $\mathbf{A} \rightarrow \mathbf{A}_{EM}$  converges relatively quickly for all starting values: on the order of 600 iterations, or half a minute, for  $n = 1000$  (given our stopping criteria). Large  $A_{EM,k}$  values typically emerge after two or three iterations, and most change, absolutely speaking, occurs in the first fifty to one hundred iterations. For error estimation, we can provide values for the minimum error possible through an inversion of the Fisher information matrix (see Appendix A.3 for brief derivation). Although we have an idea of what the best possible errors are, such values exclude the use of more standard approaches to assessments of parameter estimation, like the reduced  $\chi^2$  statistic.

*Derivation of the minimum error on EM estimates*

The asymptotic covariance matrix of  $\hat{\mathbf{A}}_{EM}$  can be approximated by the inverse of the observed Fisher information matrix,  $I$ .

As  $A_{EM,m} = 1 - \sum_{j=1}^{(m-1)} A_{EM,j}$ , there are only  $m - 1$  free parameters. Thus let  $\mathbf{A}'_{EM} = (A_{EM,1}, \dots, A_{EM,(m-1)})$ . Using  $f_{ij} = f_j(x_i, y_i)$  for brevity, the likelihood can then be expressed as:

$$\ell(\mathbf{A}'_{EM}) = \sum_{i=1}^n \log \left\{ \left( \sum_{j=1}^{m-1} A_{EM,j} f_{ij} \right) + (1 - A_{EM,1}, \dots, A_{EM,(m-1)}) f_{im} \right\} \quad (\text{A6})$$

The observed information matrix,  $I$ , is the  $(m-1) \times (m-1)$  negative hessian of Eqn. A6, evaluated at the observed data points:

$$I(\mathbf{A}'_{EM} | \mathbf{x}, \mathbf{y}) = - \frac{\partial^2 \ell(\mathbf{A}'_{EM})}{\partial \mathbf{A}'_{EM} \partial \mathbf{A}'_{EM}^T} = - \begin{bmatrix} \frac{\partial^2 \ell(\mathbf{A}'_{EM})}{\partial^2 A_{EM,1}} & \frac{\partial^2 \ell(\mathbf{A}'_{EM})}{\partial A_1 \partial A_2} & \cdots & \frac{\partial^2 \ell(\mathbf{A}'_{EM})}{\partial A_1 \partial A_{(m-1)}} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 \ell(\mathbf{A}'_{EM})}{\partial A_{(m-1)} \partial A_1} & \frac{\partial^2 \ell(\mathbf{A}'_{EM})}{\partial A_{(m-1)} \partial A_2} & \cdots & \frac{\partial^2 \ell(\mathbf{A}'_{EM})}{\partial^2 A_{EM,(m-1)}} \end{bmatrix}$$

where

$$\frac{\partial \ell(\mathbf{A}'_{EM})}{\partial A_{EM,k}} = \sum_{i=1}^n \frac{f_{ik} - f_{im}}{\sum_{j=1}^m A_{EM,j} f_{ij}} \quad \text{and} \quad \frac{\partial^2 \ell(\mathbf{A}'_{EM})}{\partial A_{EM,k} \partial A_{EM,r}} = - \sum_{i=1}^n \frac{(f_{ik} - f_{im})(f_{ir} - f_{im})}{(\sum_{j=1}^m A_{EM,j} f_{ij})^2}$$

with  $1 \leq r \leq m-1$  such that  $(k,r)$  represents the index of the observed information matrix  $I$ .

The observed information matrix of  $\mathbf{A}'_{EM}$  yields the following estimates for covariance and correlation for all  $m$  estimated weights in  $\hat{\mathbf{A}}_{EM}$ :

$$\text{Cov}(\hat{A}_{EM,p}, \hat{A}_{EM,q}) = \begin{cases} [I^{-1}(\hat{\mathbf{A}}'_{EM})]_{pq} & p, q < m \\ - \sum_{j=1}^{m-1} \text{Cov}(\hat{A}_{EM,j}, \hat{A}_{EM,q}) & p = m, q < m \\ \sum_{j=1}^{m-1} \sum_{k=1}^{m-1} \text{Cov}(\hat{A}_{EM,j}, \hat{A}_{EM,q}) & p, q = m \end{cases}$$

$$\text{Var}(\hat{A}_{EM,j}) = \sigma_j^2 = \left\{ \text{Cov}(\hat{\mathbf{A}}_{EM}) \right\}_{jj}$$